THE UNIVERSITY OF ALBERTA

LONG TERM PROJECTIONS OF THE UTILIZATION OF CRUDE OIL IN THE UNITED STATES

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF ECONOMICS
and
DEPARTMENT OF MATHEMATICS
by
© ARTHUR FITZPATRICK

EDMONTON, ALBERTA
SPRING, 1971

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled; "Long Term Projections of the Utilization of Crude Oil in the United States" submitted by Arthur FitzPatrick B. Sc., M.A., P. Eng. in partial fulfilment of the requirements for the degree of Master of Science.

ABSTRACT

A number of models to describe the time pattern of the utilization of crude oil in the continental United States (excluding Alaska) were proposed in this study. Whenever possible, boundary conditions were also stipulated, which were intended to reflect the characteristics of growth, and eventual decline, in the rate of utilization of oil. Only two of the original twenty-two models proposed, (the generalized logistic, and a model involving the pattern of percent recovery) gave a reasonable estimate of the ultimate recoverable resource base, and were not rejected because of the tests imposed on the models and on their error terms, or residuals. While the residuals were not entirely random, the model with the minimum sum of squares for error, and which was selected as a logical candidate to project annual production was the generalized logistic curve. This model implied the ultimate resource of recoverable oil in the United States would be over 300 billion barrels, with annual production reaching a maximum in the year 1985. The upper limit of recoverable oil would probably be less than 400 billion barrels, because this was the figure estimated for the ultimate discoveries of oil in place.

ACKNOWLEDGEMENTS

This study is an outgrowth from an Enquiry commissioned by the Province of Alberta, "Gasoline Marketing in the Context of the Oil Industry", and undertaken by the Gasoline Marketing Enquiry Committee. Kenneth A. McKenzie, Q.C., was chairman of the committee; Allan N. Rose and myself were members. The Enquiry was commissioned in 1965, and the report of the committee was submitted in February, 1969 to the Honorable Mr. A. R. Patrick, Minister of Industry and Tourism for the Province of Alberta.

I found the oil industry study to be a stimulating, if not genuinely rewarding experience. My appreciation is extended to all those who so generously helped me in deriving some understanding of the industry.

This study could not have been completed without the continued guidance from Dr. J. R. McGregor, Chairman of the Department of Mathematics. I am also especially grateful to the interest and encouragement of Dr. W. D. Gainer of the Department of Economics.

The detailed and exacting task of gathering data and initial analysis was carried out mainly by Mr. Marc Bolander, and Miss Rose Bilski, who worked under a National Research Council Grant through the Department of Mathematics.

Dr. B. Von Hohenbalken, Professor of Economics, was of great assistance in the final design of the study. Over the years in which this study took shape, Professor E. S. Keeping of the Department of Mathematics offered many valuable ideas for the analysis of the models.

Arthur FitzPatrick

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

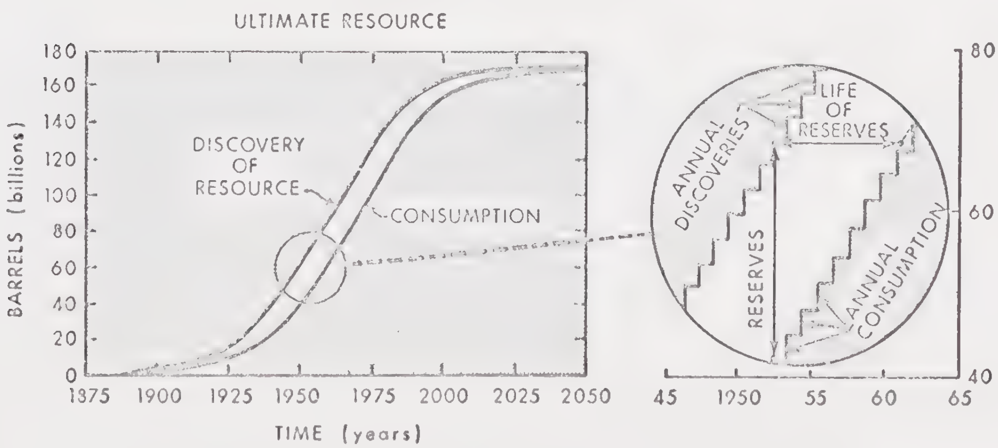
The purpose of this study is:

1. to learn what factors relate to the long term availability and utilization of oil in the United States,
2. to study and develop mathematical models for use in long term projections of the rate of discovery and utilization of oil,
3. to study any relationships which may exist between measures of the availability and utilization of oil, and
4. to apply the models in an attempt to predict future availability and utilization of oil.

This study was thus an attempt to provide an overall picture of the dynamics of the oil industry, using any long term measures that were available for the period 1859 to 1968.

The United States (excluding Alaska) was chosen as the region for investigation, because of the amount and quality of data available. More than likely, the ultimate resource of oil will eventually be discovered and consumed in that region. This possibility is shown graphically in Figure 1.1.1.

FIGURE 1.1.1



CUMULATIVE DISCOVERY AND CONSUMPTION OF A FINITE RESOURCE

SOURCE: See Reference 1.

Knowledge of either the discovery or the consumption process might allow prediction of the ultimate extent of the resource of oil, even though one would be based on estimates of the resource available, and the other on consumption by actual production of the resource.

The objective was to study some of the relationships between measures of the discovery and utilization of oil, and to incorporate these measures into a model which would predict the remaining recoverable resource of oil. The model developed should have as few a priori assumptions as necessary, and would make use of only the most reliable data.

The difficulties of relating financial factors to oil industry activities, and allocating costs to the specific products produced, has been discussed elsewhere.[2] Thus, this study was concerned with factors in which financial measures were only implicit in the actual functions used. Similarly, functions in which measures were related independently of time were also of interest.

For practical reasons, the data used pertained to crude oil - the main petroleum product obtained by the oil industry. Also, only a portion of the data available has been employed, and not all interrelationships have been analyzed. To do an exhaustive study of this complex industry would require many years. In this study, some relationships between factors have been expressed in graphical form. The illustrations are meant to be suggestive, but not proof of relationships.

REFERENCE:

1. Gasoline Marketing Enquiry Committee, Gasoline Marketing in the Context of the Oil Industry, Queen's Printer, Edmonton, 1968, p. 584.

2. Ibid., Part 8, pp. 391-516.

CHAPTER II

SOME FACTORS RELATING TO THE DISCOVERY AND UTILIZATION PROCESS

SECTION ONE - STRUCTURE OF THE OIL INDUSTRY

Much of what happens in the oil industry is the outcome of four fundamental characteristics of oil. It is difficult to find; it is not found where it is mainly wanted; it is an inflammable liquid, and it has special properties. The risk involved in finding the oil in the first place, plus having to cope with undesirable location in many cases, necessitates large capital investments. Additional capital is required for transportation and refining facilities. Finally, for major firms to insure an adequate return on all capital that is invested, marketing facilities are necessary to sell the spectrum of products derived from the oil. Therefore, fully integrated oil companies are usually very large in terms of capital investment.[1]

The continued growth of the freeworld petroleum industry involved about $11.2 billion yearly for capital expenditure during the early 1960's, and is expected to rise to about $20 billion yearly by 1975 - giving a total of about $266 billion dollars during the 1960 - 75 period. Of this, 60% is to be allocated to the replacement of existing capacity and 40% for growth. It is not practical for the petroleum industry to draw its funds from the stock markets, because the 40% needed for growth expenditure in oil alone represents about half of all funds raised on capital markets. In any case, the petroleum industry is mostly self-financed, having drawn only 6% of its funds from capital markets during the ten years preceding 1963.[2,3]

It is apparent that where the oil industry has grown to prominence you find bigness and vertical integration. The largest international oil companies

are concerned with maintaining their own positions in the world oil business. They look to permanence; they are in business to stay, and ideally to stay with a big share.

The oil companies have, in general, remained oil companies, their profits being either distributed or ploughed back into one aspect or another of the oil business. Few have put their profits into other enterprises, thus facilitating expansion by vertical integration. Eight major international oil companies are presently responsible for over 80% of crude oil production, 71% of refining capacity, 35% of tanker ownership and approximately 70% of the distribution and marketing of oil products.[4]

In 1965, this major international group of companies produced 4,761,790,000 barrels of oil in countries with net exports. If one assumes an average profit of 50 to 65 cents per barrel on this oil, operations in these exporting countries accounted for 65% to 84% of the group's total profits of 3,665 million dollars during 1965.[5,6]

Outside the United States, Mexico, and Russia, the operations of these major oil companies are combined through various inter-company holdings in subsidiary and affiliated companies. These holdings constitute partnerships in various areas of the world. Each of these companies has pyramids of subsidiary and affiliated companies in which ownership obviously provides opportunity, and even necessity, for joint action. Cooperation is achieved through such indirect means as interlocking directorates, joint ownerships of affiliates, intercompany crude purchase contracts, unitized production arrangements and marketing agreements.

Governments have, increasingly, begun to participate directly in oil operations. Countries in which government participation in oil is a complete monopoly include all the communist world, as well as Mexico, Brazil, Chile, and Taiwan. However, many other countries hold monopolies in certain sectors of their economy through various state enterprises. Thus, Spain monopolizes marketing while it admits private-enterprise companies to exploration, and Italy holds a virtual monopoly over natural gas production through ENI. In all, a total of

68 national oil companies were operated by various governments around the world at the beginning of 1965. Governments have elected to cooperate with private-enterprise capital in arrangements which are mutually beneficial.

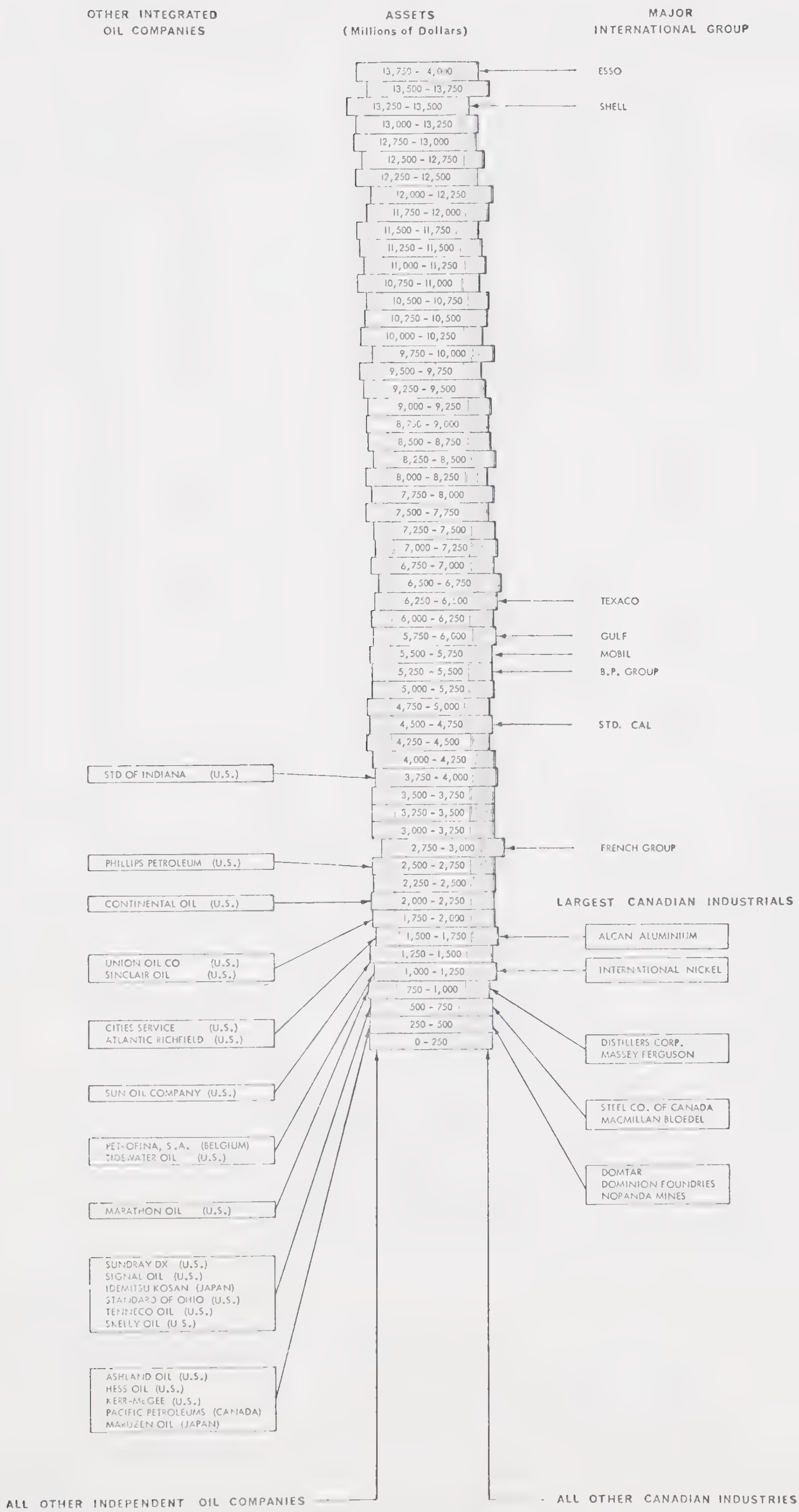
Governments have also elected to form partnerships among themselves. An example of this was the formation of Organization of Petroleum Exporting Countries (OPEC) in 1961. This organization is dedicated to maintenance of the level of posted prices. Crude oil prices are of major importance to these governments, since their tax and other revenues from oil depend on price.[7]

Thus, at each level of the industry, there are a number of activities carried out, each of which probably relates in some way to the rate of utilization of domestic oil resources for any region. The fact that a few major oil companies have a significant share of production and market may cause interaction in the activities between regions. In any case, shut-in production for one region has always been compensated by accelerated production elsewhere.

An ever quickening pace, however, has generally characterized the oil industry, regardless of reason. Per capita demand for petroleum has increased in major consuming countries, increase in population compounds the increase in demand. More and more producing areas thus have become established to meet the need. Initial production, while low, usually increases to some maximum for the area and then declines after the area is fully explored and developed. This has been the case in a number of European countries, and for some areas in the United States.[8,9]

FIGURE 2. 1. 1

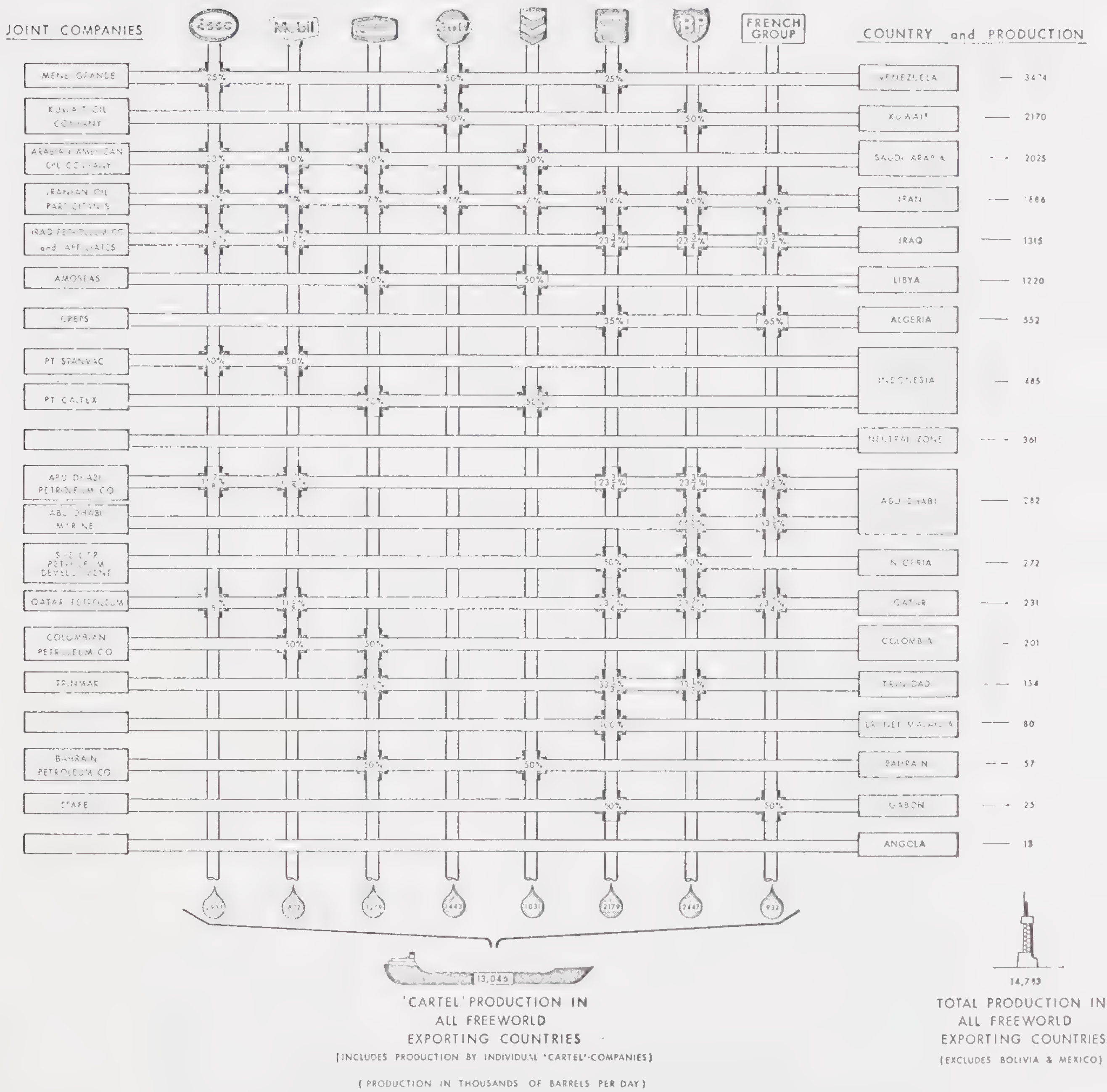
ASSETS OF ALL MAJOR OIL COMPANIES COMPARED
WITH CANADIAN INDUSTRIAL FIRMS
1966



SOURCE: See Reference 10

FIGURE 2. 1. 2

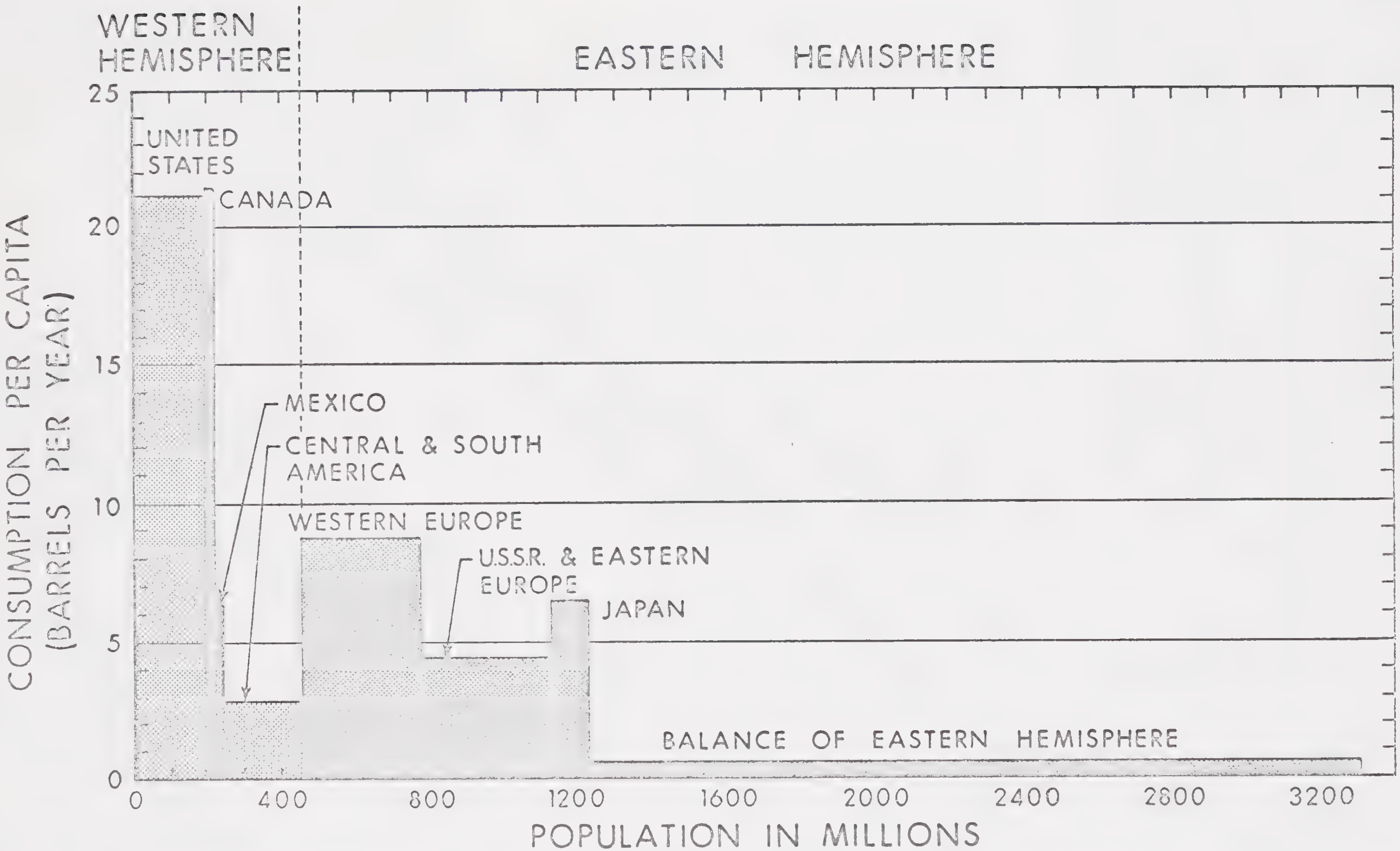
'CARTEL' SHARE OF PRODUCTION AND/OR OWNERSHIP OF EXPORTED OIL
1965



SOURCE: See Reference 11.

FIGURE 2. 1. 3

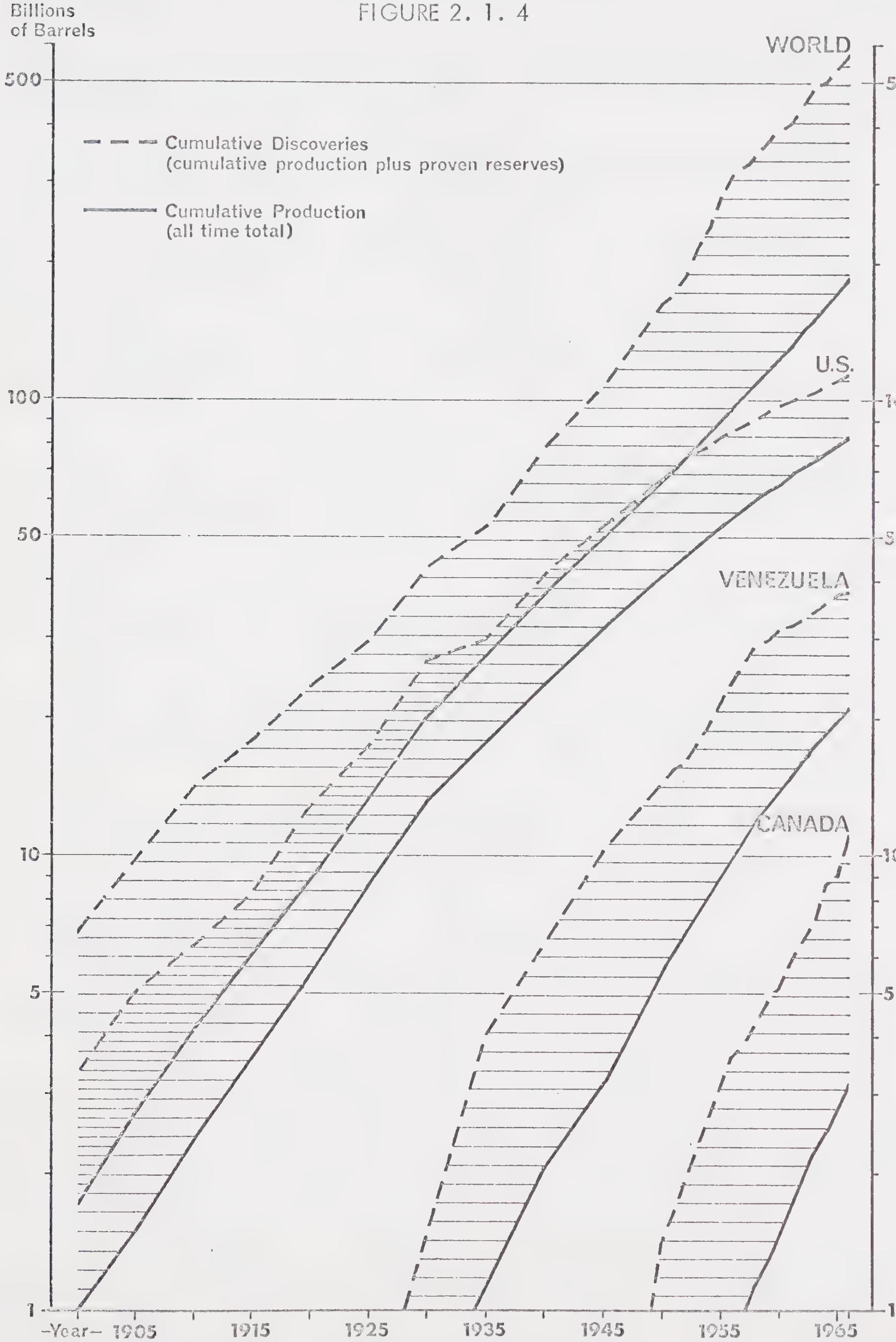
PER CAPITA CONSUMPTION OF OIL, 1965



SOURCE: See Reference 12.

CUMULATIVE PRODUCTION & CUMULATIVE DISCOVERIES OF OIL 1900 – 1966

FIGURE 2. 1. 4



SOURCE: See Reference 13.

REFERENCES:

1. Shell International Petroleum Company Limited, The Petroleum Handbook, Shell Centre, London SE1, 1966, p.3.

2. G. H. Barrows, International Petroleum Industry, Volume I, International Petroleum Institute Inc., New York, 1965, p. 13.

3. P. R. Odell, An Economic Geography of Oil, G. Bell and Sons, Ltd., London, 1963, p. 28.

4. U. S. Federal Trade Commission, The International Petroleum Cartel, Washington, 1952, pp. 23-29.

5. See Figure 2.1.2.

6. Gasoline Marketing Enquiry Committee, Gasoline Marketing in the Context of the Oil Industry, Queen's Printer, Edmonton, 1968, pp. 396-401.

7. Barrows, pp. 35-36.

8. H. V. Strugh, Editor, The Petroleum Data Book, The Petroleum Engineer Publishing Co., Dallas, Texas, 1947, pp. E14-E85.

9. Degolyer and McNaughton, Twentieth Century Petroleum Statistics, Degolyer and McNaughton, Dallas, Texas, Annual, 1968, pp. 3 - 14, and pp. 18 - 24.

10. Gasoline Marketing Enquiry Committee, p. 568.

11. Ibid., p. 589.

12. Ibid., p. 580.

13. Ibid., p. 582.

SECTION TWO - OIL INDUSTRY ACTIVITIES

In the integrated oil industry, there have traditionally been six distinct activities in the utilization of petroleum, as follows:

1) Exploration: This is the search for pools of oil and/or gas. Exploration operations include: aerial surveys, geophysical surveys, geological studies, core testing, drilling of test wells (i.e. - wildcat wells), and the acquisition of mineral rights or permits or reservations from those holding mineral rights.

2) Development: Development is the drilling and/or bringing into production of wells or a field, following exploration activity and the discovery of a pool. Development activities include development drilling and the provision of equipment and facilities necessary for the efficient extraction, gathering, and conservation of the petroleum.

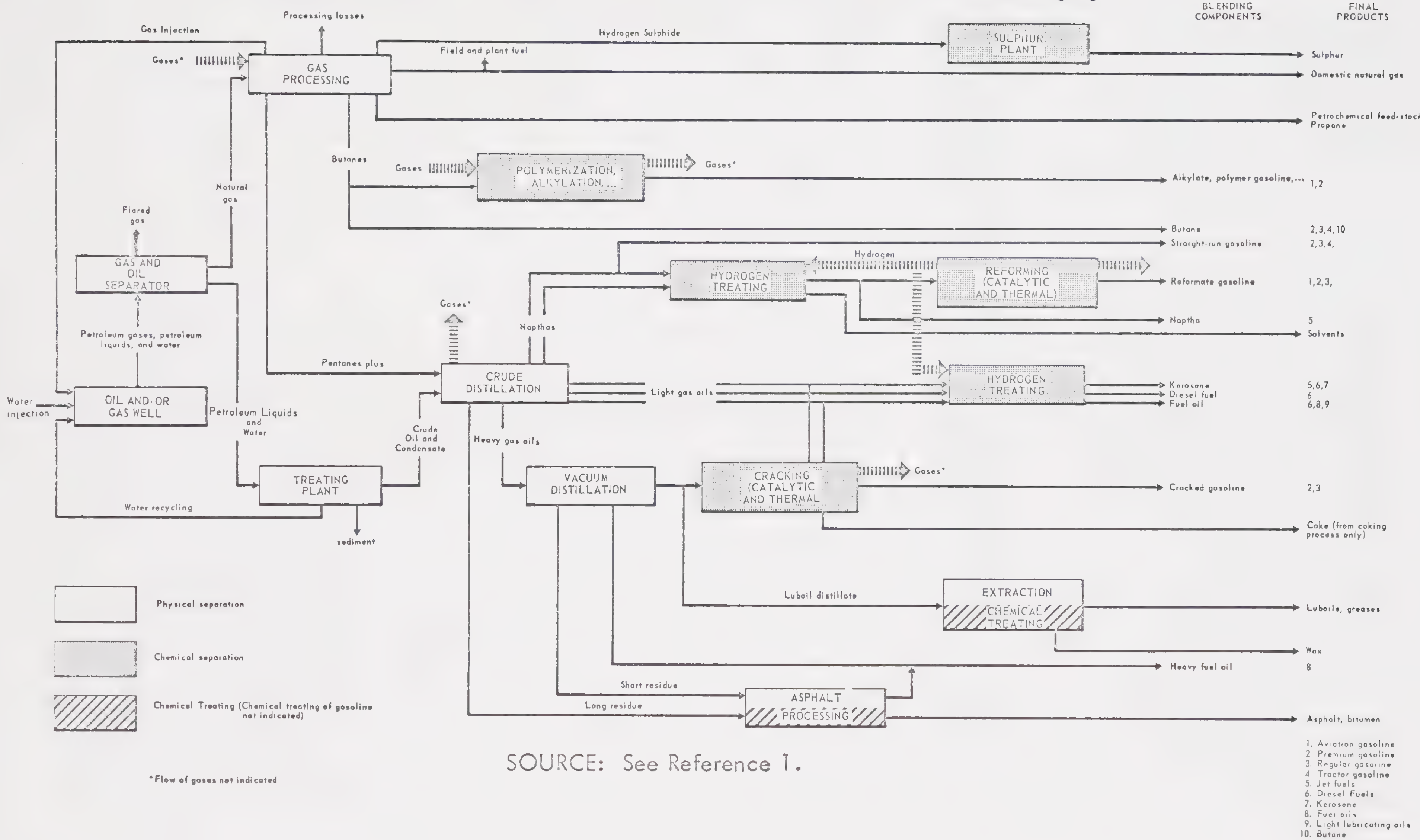
3) Production: Production is the act or process of producing oil or gas. The act of production usually involves royalty payments to the owner of the mineral rights.

4) Transportation: Transportation is the act or process of transporting petroleum products such as crude oil, natural gas, etc., to or within a market area. The distribution of refined products is not generally considered part of the transportation activity, even though some areas do have pipelines for refined products, and some are shipped by boat. However, the bulk of refining is increasingly carried out near a market area. Transportation commonly involves crossing of political boundaries, so that in addition to direct transportation costs, import duties, quotas, and other restrictions may be associated with transportation.

5) Refining: Refining is the act or process by which the physical or chemical characteristics of petroleum or petroleum products are changed, exclusive of the operation of placing petroleum in settling tanks to remove sediment and water, or passing petroleum through separators to remove gas. The refining process is employed to upgrade oil into a vast number of more

FIGURE 2. 2. 1

SIMPLIFIED FLOW OF PRODUCTS IN THE PROCESSING OF OIL AND GAS



SOURCE: See Reference 1.

valuable products; nearly all oil is refined prior to eventual consumption.

6) Marketing: Marketing is the final handling, selling and delivery of petroleum products which have been refined and adapted to the use of the consumer.

Some of the steps of separating and treating petroleum products into their finished form have been illustrated in Figure 2.2.1.

REFERENCE:

1. Gasoline Marketing Enquiry Committee, Gasoline Marketing in the Context of the Oil Industry, Queen's Printer, Edmonton, Alberta, 1968, p. 421.

SECTION THREE - SOME FACTORS RELATING TO THE DISCOVERY AND UTILIZATION OF OIL.

The dynamic relationship between some factors (each of which describes in part the process of discovery and utilization of oil) has been charted in the simplified flow diagram in Figure 2.3.1.

The initial boom in the industry continues to be manifest, but to a diminishing extent, as the resource is depleted. The most accessible sedimentary basins are naturally among the first to be explored. In these basins, the largest or most accessible fields are more readily found, and thus brought into production first. While these larger fields are being gradually depleted, somewhat smaller (and/or more remote) fields tend to be discovered; thus, this process is repeated over a long period of time.[1]

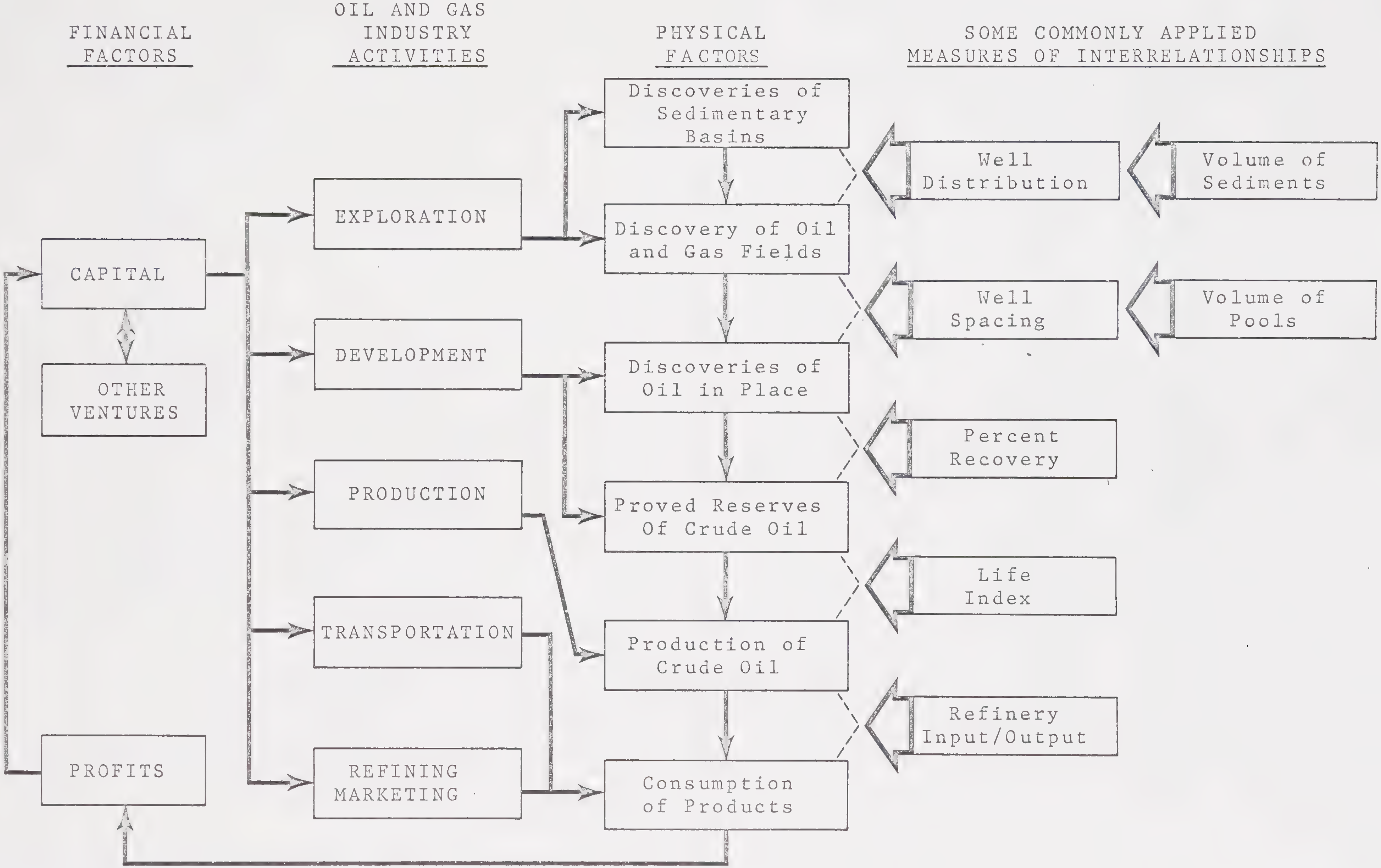
Improved secondary recovery, and the use of drilled facilities which may have been financially depreciated to the write-off stage, tend to extend the life of many fields well beyond their original expectancy.[2] However, the amount of effort necessary to produce the entire recoverable resource may increase as the last barrels of oil are produced. As time goes on, much of the capital accumulated will either be used to find and produce oil in other areas, or be applied to other ventures.

Towards the end of the life of a field, each new barrel of proved reserve would be produced almost as soon as it is found. If current production costs alone become greater than the value of the oil, physical production may cease.

Sporadic efforts to apply new techniques for secondary recovery would draw out more oil, but at an ever diminishing rate - until that time when all the resource was used. The life cycle for other sedimentary basins and for fields within these basins would be parallel - unless they were never discovered.

The strong economic forces, which in the beginning, have activated and nourished the growth in production, would no longer be great enough to sustain further production. All the accumulated wealth, and economic resources, would be subject to some law of diminishing returns, and so would be directed to other ventures.

FIGURE 2.3.1
RELATIONSHIP BETWEEN SOME FACTORS IN THE DISCOVERY AND UTILIZATION OF OIL



REFERENCES:

1. J. J. Arps and T. G. Roberts, Economics of Drilling for Cretaceous Oil on East Flank of Denver - Julesburg Basin, Bulletin of the American Association of Petroleum Geologists, Volume 42, Number 11, November, 1958, p. 2549 - 2566.

2. E. B. Miller, Jr., "Old Oil Fields Never Die", Economics of the Petroleum Industry, Volume 3, Gulf Publishing Company, Houston, 1965, pp. 137 - 150. (A postscript to the title might be "They Just Drain Away").

CHAPTER III
RESULTS OF PREVIOUS STUDIES

SECTION ONE - INTRODUCTION

Numerous estimates of the ultimate resource of crude oil in particular regions have been made. While most have been for the United States, a few have included other major producing countries.

Perhaps the most widely known studies have been carried out by Lewis G. Weeks, who was a petroleum geologist for Standard Oil of New Jersey. Weeks used all information which was available, such as statistics about production, reserves and wells drilled, geophysical readings, along with any other data accessible, as an aid in arriving at an estimate of the volume of sediments and the ultimate recoverable resource of oil in all known producing areas of the world.

Figures 3.1.1 to 3.1.6 are graphical summaries of some estimates which have been made of the world resources of fossil fuels. The estimates of world crude oil resources were based mainly on the studies by Weeks.[1,2]

The results of such studies, which are sometimes based on knowledge of a variety of information, are of great interest. Presumably a properly devised dynamic model should lead to similar results. To the extent that an adequate amount of reliable data is included in the model, a projection should be obtained which may in time prove to be fairly valid.

According to C.L. Moore, "The projected patterns of petroleum industry activities .... are the conceptual patterns of these future activities. It is certain that these projected patterns do not accurately portray such future activities. Rather, they provide a reference base or guide from which to judge the adequacy with which such projected activities will meet the projected requirements for domestic petroleum, ...

" The signficance and value of projections of future activities depends on their probable reliability. The reliability of any projections of historic patterns into the future must be assessed from two completely different approaches -- the

conceptual and the statistical. If the concept of continuity is accepted, i.e. activities are valid criteria of the future patterns of such activities for a useful period of time, then the only question is the reliability of the method used for analyzing and projecting the historic patterns. Conversely, if this basic concept of continuity is rejected, then all methods for projecting historic patterns are useless. Acceptance of the opposite concept implies that the future pattern of established activities cannot be predicted.

"If the concept of continuity in the historic pattern of relevant activities in the petroleum industry is accepted, then the reliability of the method used for the analysis and projection of such patterns can be appraised by established statistical criteria and techniques."[3]

In the process of utilization of the resource, several indications of the probable course of events may emerge. Unfortunately, in the early stages of utilization, there is sometimes little on which to base any sound projection, and while all data are probably relevant, some indicators are probably much more valid than others.

Measures of production, proved reserves, oil in place, and wells drilled, form the basis for projections that have been made concerning the ultimate resource of crude oil. When the latter two measures alone are used as a basis, problems arise in the estimation of the economics, or timing of utilization. On the other hand, when production, or proved reserves are used as a basis, the projections obtained may differ unrealistically from geological estimates of the amount of oil in the ground. No study has related the measures to each other in an analytically consistent model with meaningful cause and effect relationships.

The geological studies tend to ignore economic and political factors. And because of the relatively short-term perspective and fairly urgent nature of activities associated with proving reserves and producing them, projections based strictly on production and proved reserve data may lead to grossly differing results, - according to the kind of model used.

In order to commence an analysis of the discovery, production and utilization of oil, most studies have made several simplifying assumptions. Interactions between regions were usually not included in the analysis, and it was assumed that in a particular region, trends in discoveries and production followed a unique pattern which could be determined for that area.

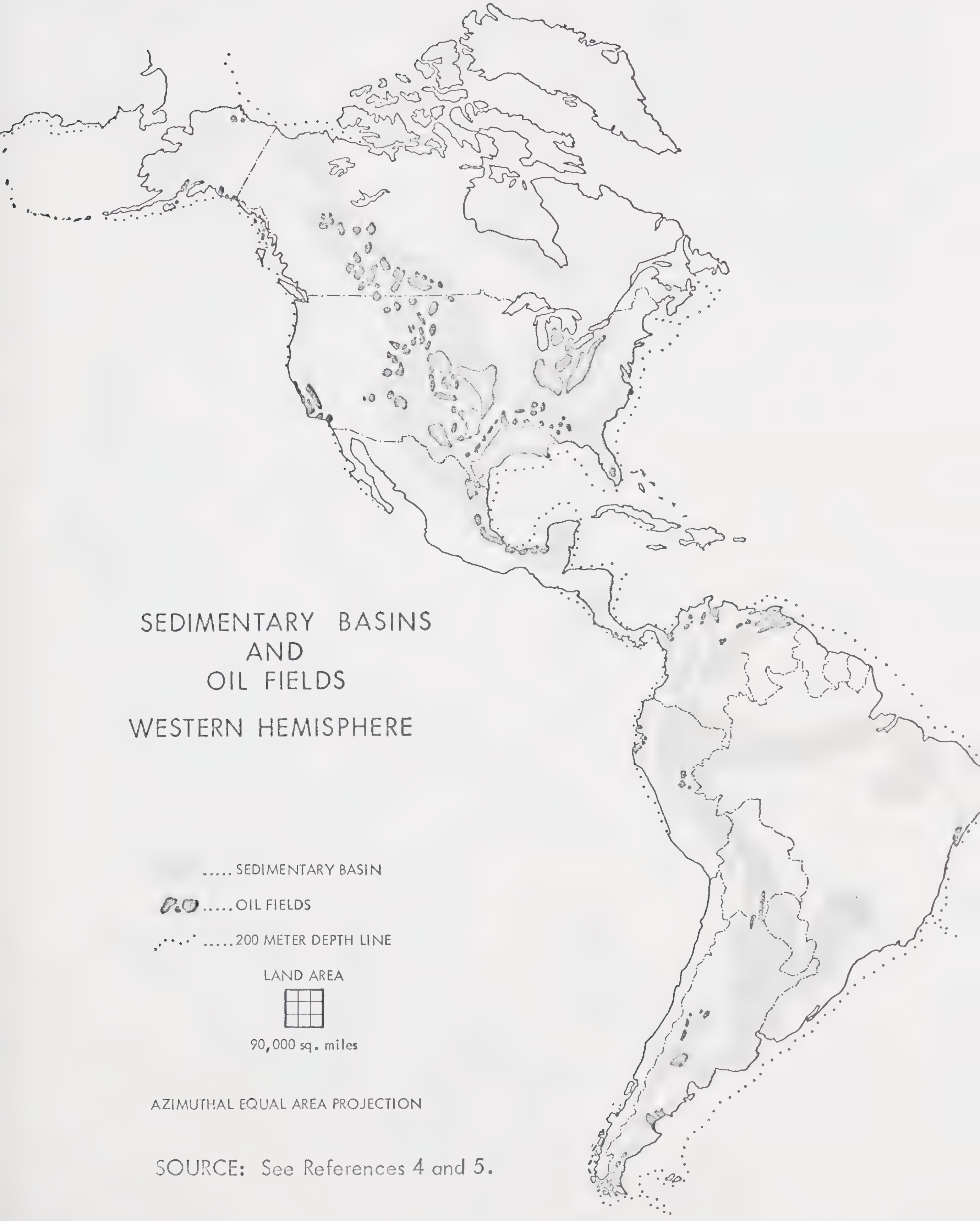
In general, it was assumed that the ultimate resource of the i^{th} region, R_i, could be determined from a projection of the time trend in discoveries, production, and/or other factors, $x_{ij}(t)$, so that as $t \to \infty$,

R_i was implied by some function of $(x_{i1}, x_{i2}, x_{i3}, \ldots x_{ij}, t)$

where j = the number of factors for which historical data were available.

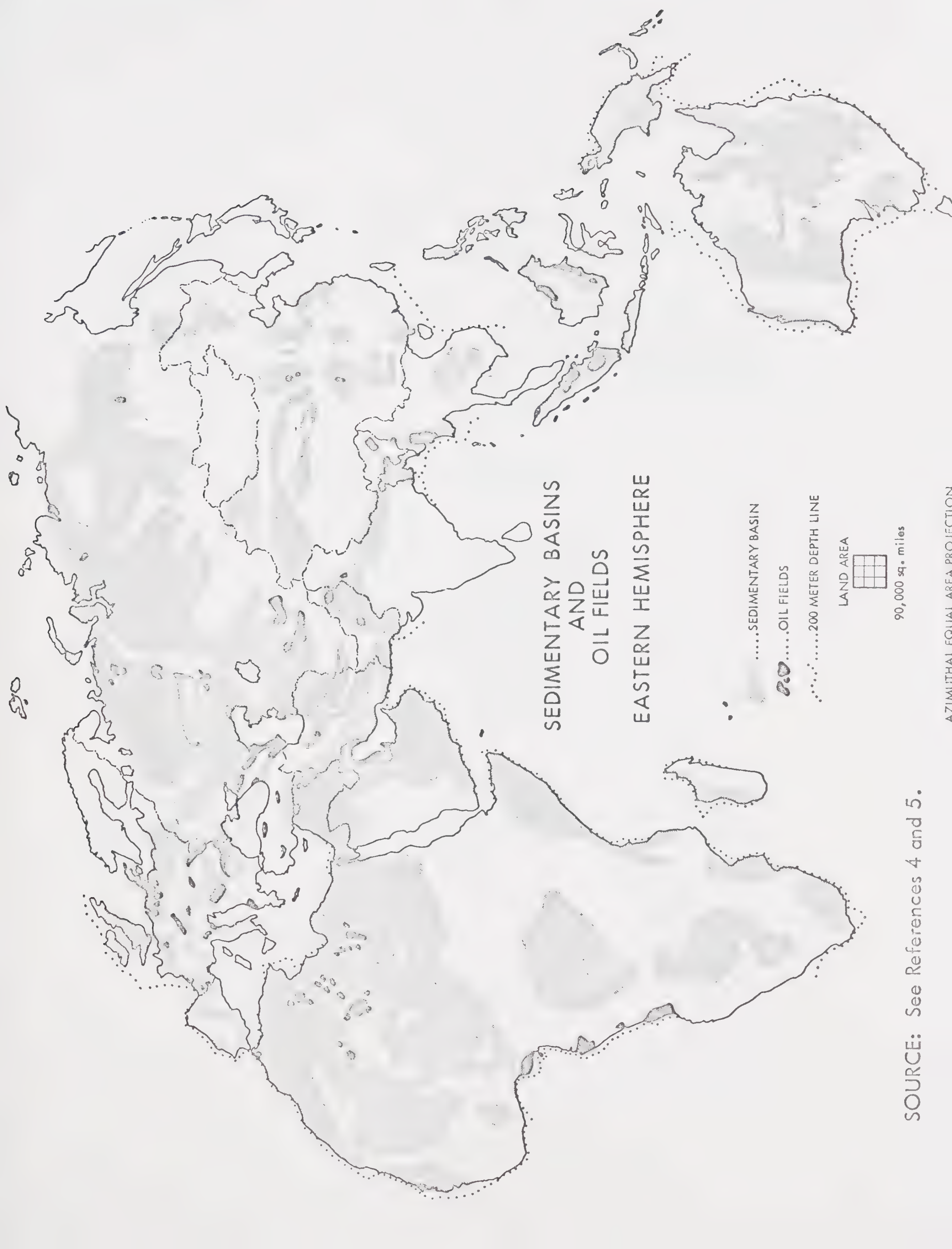
If, for some factor, reliable data were available encompassing a fairly long period of time, it was sometimes possible to use these data in a special case of the equation above. Some factors which met that requirement are discussed below, along with a summary of studies dealing with any projections of the data to the ultimate recoverable resource.

FIGURE 3. 1. 1



SOURCE: See References 4 and 5.

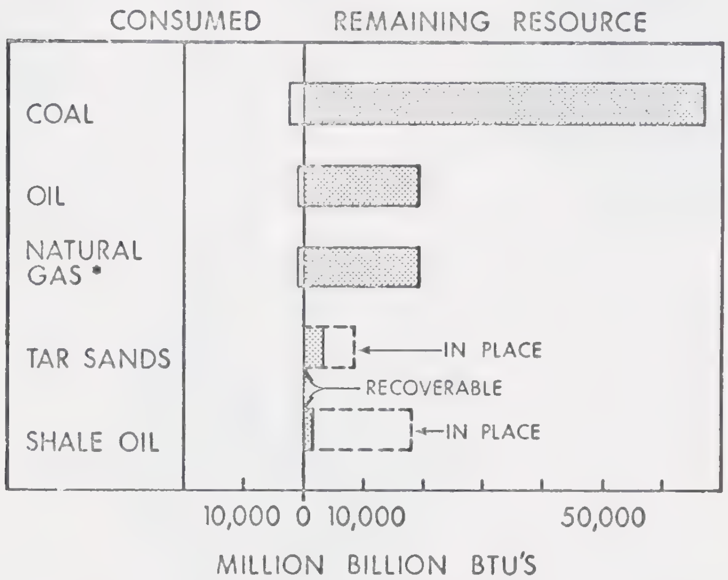
FIGURE 3. 1. 2



SOURCE: See References 4 and 5.

FIGURE 3. 1. 3

TOTAL WORLD ENERGY RESOURCES OF FOSSIL FUELS



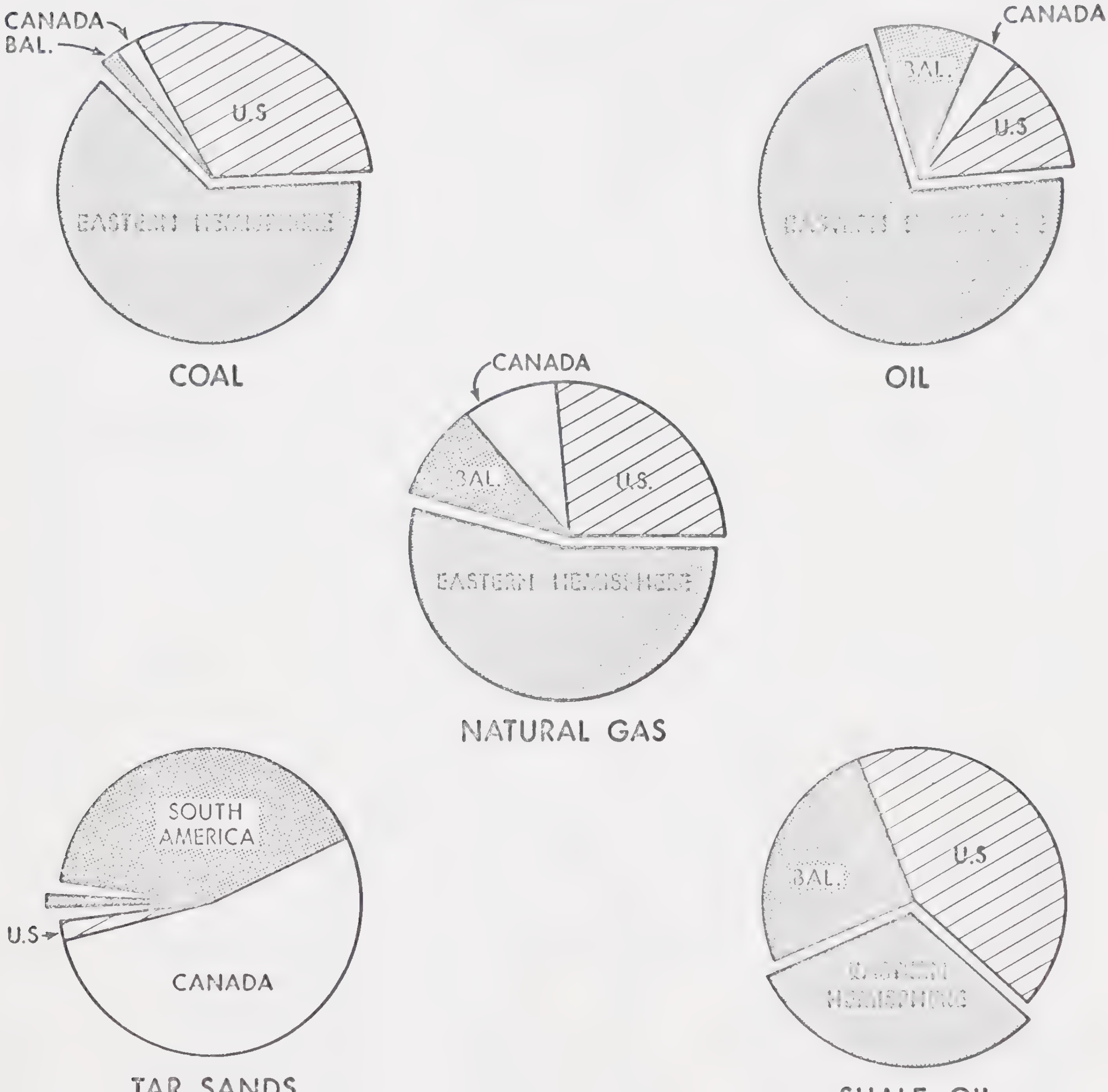
* ESTIMATED EQUAL TO OIL (THIS IS EQUIVALENT TO ASSUMING 5500 CUBIC FEET OF GAS PER BARREL OF OIL.)

The remaining resource was based on various estimates which had been made during the period 1961 to 1968; the amount consumed was as of the end of 1965.

SOURCE: See Reference 6.

FIGURE 3. 1. 4

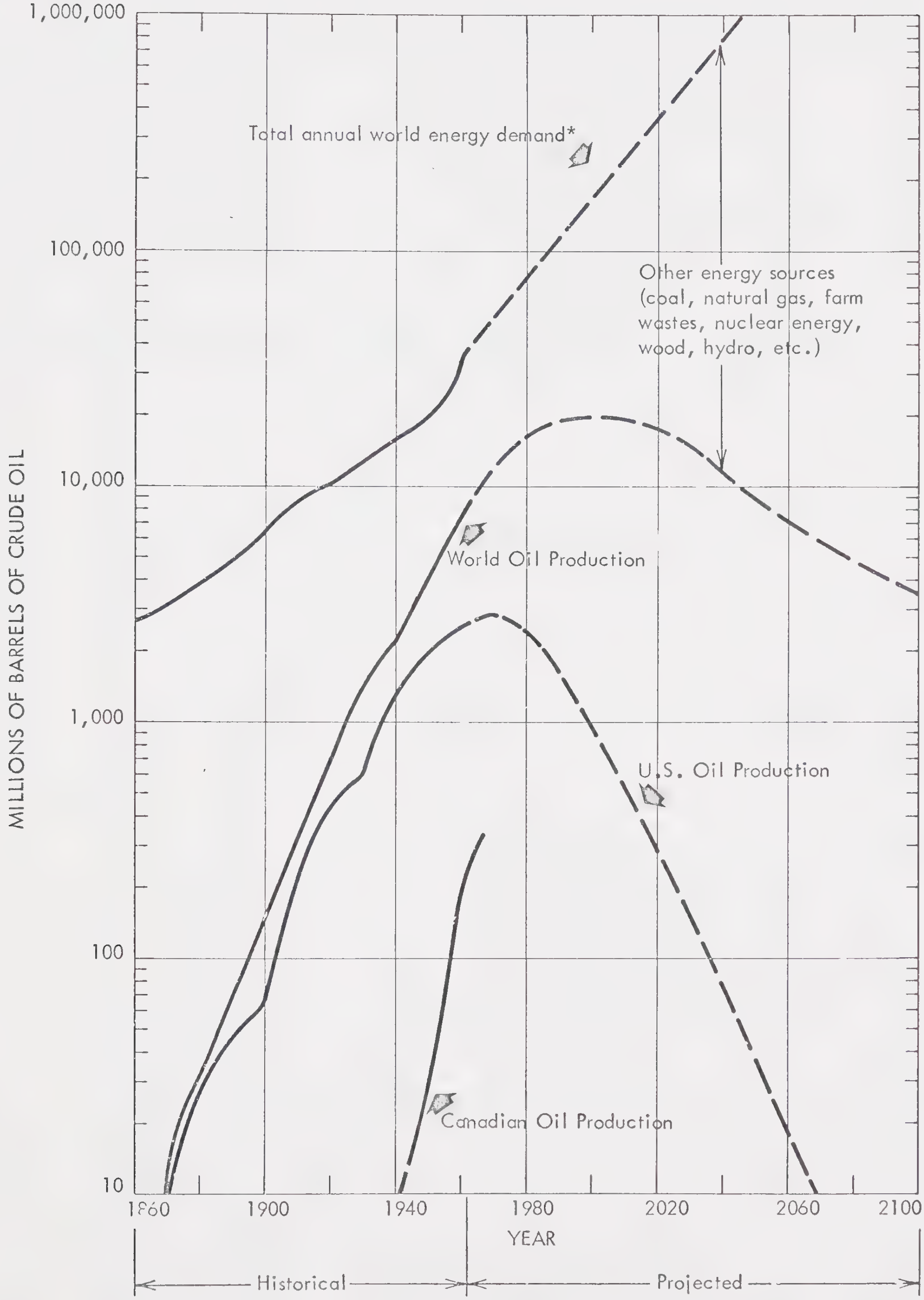
DISTRIBUTION OF WORLD ENERGY RESOURCES OF FOSSIL FUELS



SOURCE: See Reference 7.

FIGURE 3. 1. 5

PRODUCTION OF CRUDE OIL
COMPARED TO
TOTAL ANNUAL ENERGY DEMAND

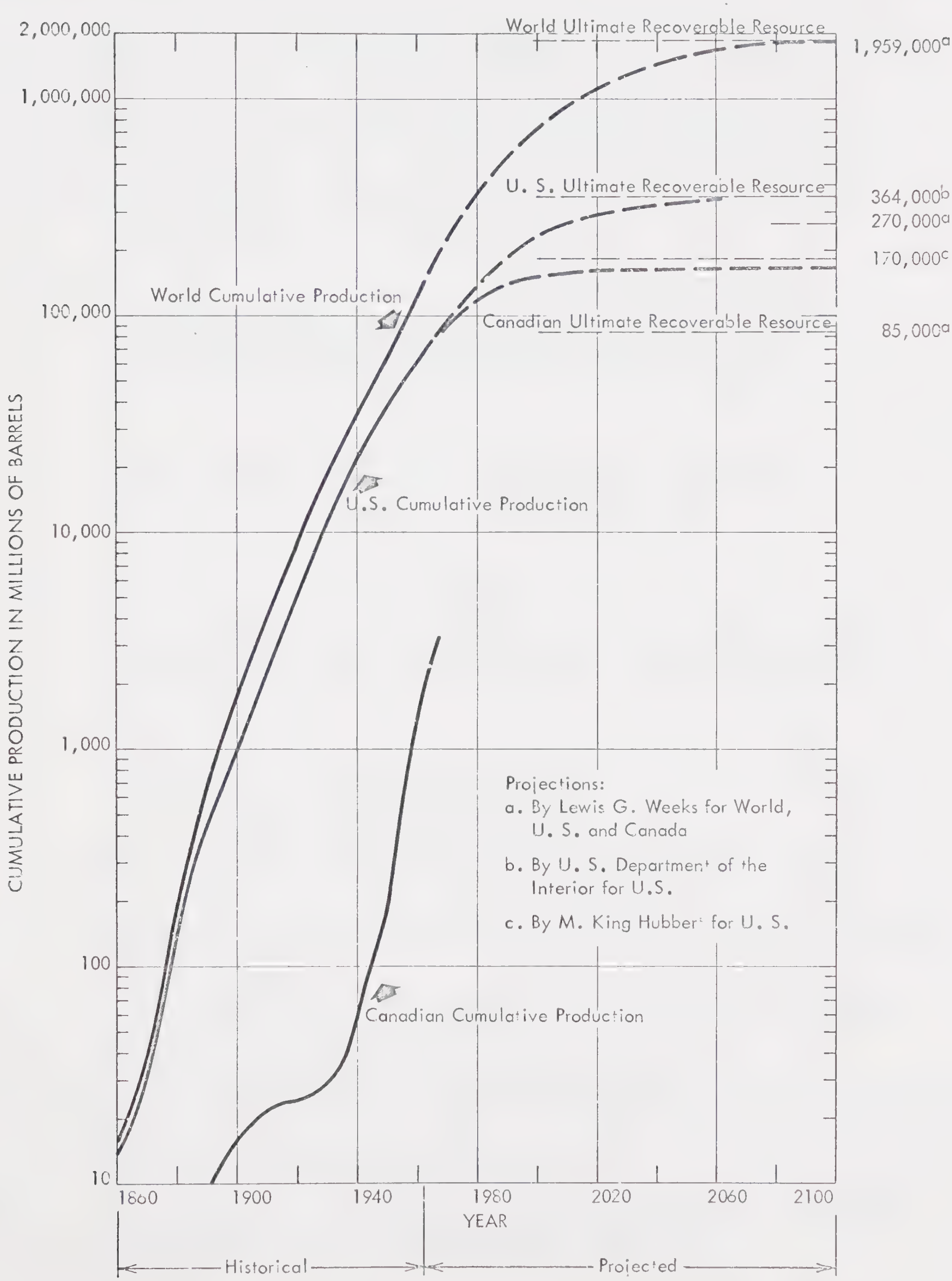


*Total annual energy demand converted to millions of barrels of crude oil.

SOURCE: See References 8, 9, 10, and 11.

FIGURE 3. 1. 6

CUMULATIVE (ALL-TIME) PRODUCTION
OF CRUDE OIL
AND PROJECTIONS



SOURCE: See References 12, 13 and 14.

REFERENCES:

1. Lewis G. Weeks, Sixth World Petroleum Congress, Proceedings, Frankfurt, June, 1963, p. 271.

2. Lewis G. Weeks, "Worldwide Review of Petroleum Exploration", Proceedings, 38th Annual Fall Meeting of the Society of Petroleum Engineers of AIME, 1963, and reprinted in G. H. Barrows, International Petroleum Industry, Volume I, International Petroleum Institute Inc., New York, 1965, pp. 283 - 285.

3. C. L. Moore, Projections of U. S. Petroleum Supply to 1980 with Annex Entitled the Gompertz Curve for Analyzing and Projecting the Historic Supply Patterns of Exhaustible Natural Resources, U. S. Department of the Interior, Office of Oil and Gas, Washington, D. C., 1966, p. 1.

4. American Geographical Society, World Geography of Petroleum, W. E. Pratt and D. Good, Editors, Princeton University Press, 1950, Map: "World Sedimentary Basins and Petroliferous Areas".

5. Lewis G. Weeks, Worldwide Review..., pp. 283 - 285.

6. Gasoline Marketing Enquiry Committee, Gasoline Marketing in the Context of the Oil Industry, Queen's Printer, Edmonton, Alberta, 1968, p. 584.

7. Ibid., p. 583.

8. G. H. Barrows, International Petroleum Industry, Volume 1, International Petroleum Institute Inc., New York, 1965, pp. 9, 25.

9. See Figure "Cumulative Production and Cumulative Discoveries of Oil".

10. M. King Hubbert, Energy Resources, National Research Council, Washington, D. C., 1962, p. 59.

11. P. S. Putnam, Energy in the Future, D. Van Nostrand Co. Inc., Princeton, N. J., 1953, pp. 148, 326, 439, 440, 441, 442.

12. See Figure "Production of Crude Oil Compared to Total Annual Energy Demand".

13. Barrows, International Petroleum Industry, p. 9.

14. U. S. Department of the Interior, Plain Facts About Oil, U. S. Department of the Interior, Washington, D. C., 1964, p. 34.

SECTION TWO - PROJECTIONS OF CUMULATIVE PRODUCTION

In a study for the United States Department of the Interior, C. L. Moore used the Gompertz curve to project cumulative production, y. Cumulative production was therefore assumed to be described by the equation

$$y = Rc^{d^t} \tag{3.2.1}$$

where c and d were constants, with $\ln|\ln c|$ establishing the time origin and R the upper asymptote of y.[1,2]

The differential equation from which the Gompertz equation can be derived is

$$\dot{y} = \frac{dy}{dt} = -k\, y \ln (y/R),$$

$$\text{with} \quad k = -\ln d, \text{ a rate parameter.} \tag{3.2.2}$$

When plotted against time, this growth curve has an early inflection point, and a long tail to the right, with $\dot{y}$ approaching zero as y approaches R. In 1966, Moore used data for oil production from 1920 - 1965, and estimated the ultimate resource of crude oil in the United States to be 434 billion barrels. (Actually, Moore projected rates only to 1980; R was one of the parameters determined in this projection.)[3]

Moore's method of fitting the Gompertz curve to the data was to plot Δ log y against log y, where Δlog y as defined as the difference between values of log y for successive years. The constants for Equation 3.2.2 were then obtained from the regression of Δlog y on log y.[4]

The result obtained using production data was consistent with estimates of the cumulative discoveries of crude oil in place in the United States. These were estimated at 330 billion barrels as of the end of 1960, and 381 billion barrels at the end of 1966.[5,6] Projections of estimates of crude oil in place are dealt with later in this chapter.

Moore pointed out that cumulative discoveries are calculated from:

a) estimates of proved reserves, and

b) data on actual production.

He, and others, considered the facts about actual production superior to the figures for reserves in any derivation of resource estimates.[7,8]

Moore states that "The primary criterion (for a measure of the historic pattern of activities in the oil industry) is the adequacy and accuracy of the basic statistical data used. In recent years, a few significant economic studies by competent analysts have been published in which the many variables affecting domestic petroleum supply have been related in mathematical form and by statistical techniques. As the authors of these papers have stated, their econometric studies have been severely handicapped by the lack of basic data both in quantity and quality".[9]

Annual data on domestic petroleum production in the United States were available for any period from 1859 to the present; data on proved reserves exists only from 1920 onward.

Similar information concerning activities such as expenditures on exploration is not available, and even if it were, there would be a problem in determining the proportion of expenditures for oil and the proportion of expenditures for gas.

Also, reliable information on imports and exports is difficult to assemble, because much of the crude oil is partially (or even completely) processed into a variety of products, any of which may be exported (or imported) in substantial quantities. Some indication of this flow of products from well head to consumer is shown in Figure 2. 2. 1 .

REFERENCES:

1. J. C. Ahuja, Statistical Problems in the Study of Growth, umpublished Ph.D. thesis, Department of Mathematics, University of British Columbia, June, 1963, p. 1.

2. C. L. Moore, Projections of U.S. Petroleum Supply to 1980, With Annex Entitled the Gompertz Curve for Analyzing and Projecting the Historic Supply Patterns of Exhaustible Natural Resources, U.S. Department of the Interior, Office of Oil and Gas, Washington, D. C., December, 1966, Annex, p. 5.

3. Ibid., p. 12, Annex.

4. Ibid., pp. 5 - 6.

5. C. L. Moore, Method for Evaluating U.S. Crude Oil Resources and Projecting Crude Oil Availability, U.S. Department of the Interior, Office of Oil and Gas, Washington 25, D. C., May, 1962, p. 97.

6. American Petroleum Institute, Reserves of Crude Oil, Natural Gas Liquids, and Natural Gas in the United States and Canada as of Dec. 31, 1966, Vol. 21, July, 1967, American Petroleum Institute, Washington, D.C., p. 27.

7. Wallace F. Lovejoy and Paul T. Homan, "An Alternate Approach: Calculating Reserves from Producing Capacity: The A.D. Zapp Study", Methods of Estimating Reserves of Crude Oil, Natural Gas, and Natural Gas Liquids, John Hopkins, Baltimore, 1965, pp. 70 - 78.

8. C. L. Moore, Projections of .... , pp. 1 - 2.

9. Ibid., p. 1.

SECTION THREE - PROJECTIONS OF CUMULATIVE DISCOVERIES

In a study for the National Academy of Sciences, M. King Hubbert assumed that the logistic equation described the growth equation of cumulative discoveries, q, so that

$$q = \frac{q_\infty}{1 + be^{-at}} \tag{3.3.1}$$

where q_∞ was the asymptotic value to which q would tend as time, t, became large, while a and b were positive constants, with b determining the time origin.[1] By definition, q is the sum of proved reserves, r, and cumulative production, y.

The differential equation from which the logistic equation can be derived with lower asymptote zero is

$$\frac{dq}{dt} = \dot{q} = kq\,(q_\infty - q) \tag{3.3.2}$$

where $k = a/q_\infty$ is a positive constant, or rate parameter.[2]

The method used by Hubbert to determine the constants in Equation 3.3.1 was to plot $((q_\infty/q) - 1)$ as a function of time on semilogarithmic paper using various assumed values for q_∞. The value of q which resulted in the straightest line gave the solution to Equation 3.3.1. The constants were then obtained from the equation for this line.[3]

The ultimate resource of crude oil in the United States was estimated by Hubbert to be 170 - 175 billion barrels - rather conservative figures in comparison to the resource base estimated by others. The result implied there were only 76 billion barrels to be found in all future discoveries.[4]

Hubbert assumed that any proved reserves found would be produced in 10.5 years, as had been the case for some time in the United States. Thus, cumulative production could be determined by using a time lag of 10.5 years in the logistic equation.[5]

An economist for Standard Oil of New Jersey, J. M. Ryan, criticized the method used by Hubbert on the basis that various values q_{∞} could be selected according to the time span of data used. For example, there seemed to be a change in the trend line for $((q_{\infty}/q) - 1)$ during 1930 to 1940.[6]

Ryan applied Hubbert's method to the 1861 to 1961 data, which were available for production in the United States. There appeared to be a change in trend line after 1881, and again in 1931. Using production data from 1881 to 1961, the value for q_{∞} which produced the straightest trend line was 100 billion barrels. Ryan therefore contended that an allowance for changes in the trend of discovery and production had not been incorporated into the method and that a wide range of figures could be chosen for q_{∞}; therefore the technique used by Hubbert could not be established as valid.[7]

Changes in trend for cumulative discoveries and cumulative production in the United States are also apparent in Figure 2.1.4. For the period 1930 to 1940, there is no apparent change in the trend for world production. The United States case could probably be explained by the discovery of large amounts of crude oil in place in 1930, and again in 1936 to 1938.[8] The trend change in 1880 could only be explained by some very early inflection point, because continued (logarithmic) growth at the 1860 to 1880 rates would lead to very high estimates of q_{∞}.

Hubbert's reply to Ryan was that his results were consistent with the fact that cumulative discoveries had reached an inflection point in the United States between 1940 and 1950, and that proved reserves (which precede production) had probably reached their maximum by 1960.[9,10] Hubbert contended that because proved reserves last an average of 10.5 years, production was therefore directly linked to proved reserves, with cumulative discoveries reaching a maximum before cumulative production.

However, since 1900, while the average life of proved reserves has been 11.1 years, the life of proved reserves in the United States has ranged from 9 to 15 years.[11] The relationship between cumulative discoveries and cumulative production is therefore not a simple one, and probably depends upon a variety of factors.

It is sometimes possible to obtain cumulative discovery estimates classified according to the year of discovery of fields. These estimates are called proved discoveries for fields according to their year of discovery.[12]

Proved discoveries are sometimes called ultimate recoverable reserves by year of discovery, and in Canada proved discoveries are usually called proved ultimate reserves by year of discovery. Proved discoveries may also be classified by the geological type of reservoir in which they occur.[13/14]

Cumulative discoveries can be related directly to proved discoveries with:

$$q(t) = \sum_{i=N}^{t} u_i(t), \qquad (3.3.3)$$

where $q(t)$ = cumulative discoveries estimated at year t for a region (such as the United States),

$u_i(t)$ = proved discoveries for fields which were discovered during year i; $i = N, N+1, N+2, \ldots .t,$

$u_N(t)$ = proved discoveries for all fields discovered up to and including year N, and

N = the first year in the series of data reported.

From Equation 3.3.3, it is obvious that $u_i(t)$ must be equal to the cumulative production from fields discovered during year i, plus any proved reserves remaining in these fields under the recovery conditions which existed as of year t.

With proved reserves estimated for 1964 recovery conditions, Moore used the Gompertz equation to project 1920 - 1959 proved discoveries in the

United States to an ultimate value of 139 billion barrels.[15]

This rather low estimate can be explained in part by the fact that proved discovery estimates appreciate considerably with time. It can be shown that $(n+1)$ years after discovery of a field, the relative appreciation in estimates of the proved discoveries can be rather closely approximated by a function of the form

$$\frac{u(n+1) - u(n)}{u(n)} = \frac{b(1 - e^{-a})}{e^{an} - b} \qquad (3.3.4)$$

where $$u(n) = \sum_{i=j-n}^{t-1-n} \ddot{u}_i(j), \qquad (3.3.5)$$

$u(n)$ = gross proved discoveries, aggregated for all discovery years n years after discovery (sometimes called ultimate proved reserves),

$u_i(j)$= proved discoveries as estimated at the end of the jth year of production,

$n = o, 1, 2, \ldots, t-2$,

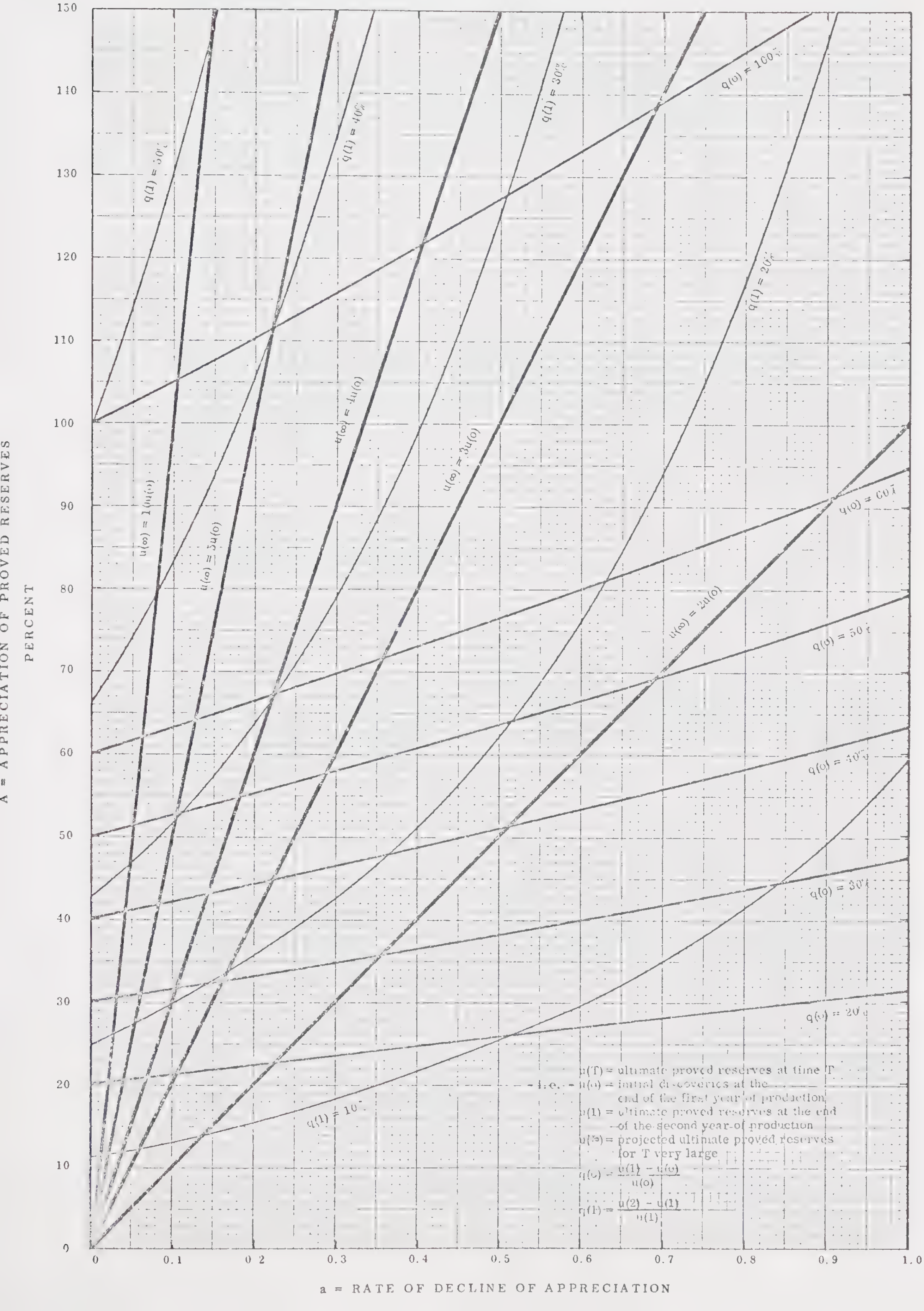
and $b = A/(a+A)$.

A and a are positive constants; A represents the appreciation of proved discoveries, and a represents the decline of appreciation in proved discoveries as proved reserve estimates are updated from year to year.[16]

The relationship of the constants A and a in Equation 3.3.4 has been shown in Figure 3.3.1. The constants can vary considerably according to the kind of proved reserve data being used; more conservative proved reserve data would lead to a larger value for A. Some typical values are $A = 1.9$ and $a = 0.1$ for proved reserve estimates in Alberta during 1948 to 1961.

FIGURE 3. 3. 1

RELATIONSHIP BETWEEN APPRECIATION OF PROVED RESERVES
RATE OF DECLINE OF APPRECIATION,
AND ULTIMATE PROVED RESERVES



REFERENCES:

1. M. King Hubbert, Energy Resources, National Research Council, Washington, D.C., 1962, pp. 57-59.

2. A. Hald, Statistical Theory with Engineering Applications, John Wiley & Sons Inc., New York, 1955, pp. 658 - 662.

3. M. King Hubbert, p. 59.

4. Ibid., pp. 51, 59, 72, 73.

5. Ibid., p. 58.

6. John M. Ryan, National Academy of Sciences Report on Energy Resources: Discussion of Limitations of Logistic Projections, Bulletin of the American Association of Petroleum Geologists, Vol. 49, No. 10, Oct. 1965, pp. 1715 - 1717.

7. Ibid., pp. 1715, 1718.

8. See Appendix A and Appendix B.

9. Ryan, pp. 1720 - 1727.

10. J. M. Ryan, "M. King Hubbert's Reply to J. M. Ryan", Limitations of Statistical Methods for Predicting Petroleum and Natural Gas Reserves and Availability, Journal of Petroleum Technology, March, 1966, pp. 285 - 286.

11. See Appendix A and Appendix B.

12. C. L. Moore, Projections of U. S. Petroleum Supply to 1980, With Annex Entitled the Gompertz Curve for Analyzing and Projecting the Historic Supply Patterns of Exhaustible Natural Resources, U. S. Department of the Interior, Office of Oil and Gas, Washington, D. C., December, 1966, Annex, p. 6.

13. American Petroleum Institute, Reserves of Crude Oil, Natural Gas Liquids, and Natural Gas in the United States and Canada as of December 31, 1968, Canadian Petroleum Association, (Joint Publisher) Calgary, Alberta, Vol 23, May 1969, pp. 104 - 106.

14. Canadian Petroleum Association, 1968 Statistical Yearbook, Canadian Petroleum Association, Calgary, Alberta. June, 1969, pp. 35, 56, 104.

15. Moore, p. 5.

16. M. King Hubbert, Degree of Advancement of Petroleum Exploration in United States, the American Association of Petroleum Geologists Bulletin, Vol 51., No. 11., November, 1967, pp. 2207 - 2214.

SECTION FOUR - PROJECTIONS OF DISCOVERIES OF OIL IN PLACE

Estimates of the amount of crude oil originally in place, according to year of discovery, are now published annually for the United States and Canada. The Canadian estimates are for discovery years from 1947 and onward, and the figures for the United States start with discovery year 1920. The cumulative total of the original crude oil in place, estimated for each discovery year up to year t, is the cumulative discovery of crude oil in place, $w(t)$. This amount is revised (usually upward) as more information is obtained about fields previously discovered. The relative amount of appreciation is usually greater for those fields most recently discovered, and is a reflection of the greater rate of appreciation in proved discoveries for a field during the first few years of production.[1,2]

In summary, if at year t, $v_i(t)$ = the estimated original amount of crude oil in place in fields discovered during year i,

$i = N, N+1, N+2, \ldots, t,$

then $$w(t) = \sum_{i=N}^{t} v_i(t) \qquad (3.4.1)$$

where N was the year production first began, and where $w(t)$ = cumulative discovery of crude oil in place, as estimated as of the end of year t.

In a study made in 1962, Moore adopted a fifteen term formula to smooth the variations in v which occurred with time. Moving averages tended to blanket the original data, and did not provide an acceptable fit with the original data; the fifteen term formula retained the inherent cyclical variations.[3]

Moore then plotted v against w, and assumed that v increased linearly with w up to 141 billion barrels (during the period 1860 to 1932) and then decreased linearly with w, as w increased to a projected terminal value of 486 billion barrels. This led to a simplified exponential formula

$$v(n) = 1700\,(e^{.062886\,n} - 1) \qquad (3.4.2)$$

for n = time in years from 1860 for the realtionship up to 1932; for the relationship after 1932

$$v(n) = 330723\ (1-e^{-.026109n}) \qquad (3.4.3)$$

Estimates of crude oil in place provide only a basis for calculating proved reserves. The next step is to project the anticipated percent recovery of crude oil originally in place.

Increases in percent recovery are probably brought about by improved technology, and by the increasing necessity to extract as much of the resource already discovered, rather than to prospect for more oil. To the extent that increases in percent recovery are a direct consequence of improved technology, percent recovery might best be described by some growth curve with time as the independent variable. The percent recovery would be low and nearly constant during the first years of production, but would likely rise asymptotically to some value less than 100%. This would likely lead to a fairly complex relationship between cumulative discoveries of oil in place and production.

As outlined in the earlier report by Moore, the second step (after projecting oil in place) was to establish the historic pattern of the percent recovery of crude oil in place.[5]

Moore used a modified logistic equation for this purpose. He assumed

$$p.r. = \frac{p.m.}{1 + 15\ e^{-at}} \qquad (3.4.4)$$

where

p.r. = cumulative percent recovery

p.m. = maximum ultimate percent recovery

t = time in years from January 1, 1930

a = a variable derived from the equation

$$\log\left(\frac{1}{a-.065}\right) = 0.52 + \frac{t}{19.931} \qquad (3.4.5)$$

Using 15 percent as the lower bound of p.r. (for the years up to 1930), Moore tried several values of p.m., and found that a value of 75 percent provided the best correlation with actual data for the years 1937 - 1961. This result seemed to Moore to be reasonable enough, and so Equation 3.4.5 was used to describe future recovery for the United States. At 75% recovery, the ultimate recoverable resource of crude oil was 364 billion barrels. [6]

The third step was to calculate projected cumulative discoveries from the results obtained in steps one and two, since

$$p.r. = \frac{q(t)}{w(t)} \quad , \qquad \text{by definition.}^{7} \qquad (3.4.6)$$

The fourth, and final step, involved the establishment of the historic pattern of proved reserves. From this, the trend in cumulative production could be obtained directly.

Moore tried various curves and formulas to match the pattern of cumulative production when plotted (linearly) against cumulative discoveries; assuming the following boundary conditions:

1. When cumulative production was zero, cumulative discoveries were also zero.
2. In the final year of production, the last barrel of proved reserves would be produced, so that both cumulative production, and cumulative discoveries, would reach 364 billion barrels, (i.e. 75% of the oil in place) and no more.

A catenary curve, which passed through the points (0, 0), (364, 364) and gave the best correlation between y and q for the series of years 1930 - 1961, was finally adopted. The pair of equations for the formula used were

$$r = b - \sqrt{2}\, a \cosh (p/a) \qquad (3.4.7)$$

$$\text{and } \sqrt{2}\, p = 366 - 2 (q-b) \qquad (3.4.8)$$

where a and b were positive constants,

p = a variable linking the two equations,

r = proved reserves,

and q = cumulative discoveries. [8]

The above three measures of physical production, proved reserves, and discoveries of oil in place, have formed the basis of most of the projections using time dependent growth functions. While there are other factors which would be relevant to the rate of utilization of crude oil, few measures are published for as great a time span as the discovery and production data. There are even fewer measures published for the major areas in which oil has been discovered and produced.

The various activities of the oil industry can be related directly to data on discoveries and production of oil. In particular, exploration experience in one area may be applied to other (less explored) areas, as discussed in the next Section.

REFERENCES:

1. Canadian Petroleum Association, 1967 Statistical Year Book, Canadian Petroleum Association, Calgary, Alberta, p. 98.

2. American Petroleum Institute, Reserves of Crude Oil, Natural Gas Liquids, and Natural Gas in the United States and Canada as of Dec. 31, 1966. American Petroleum Institute, Washington, D.C., Vol. 21, July, 1967, p. 27.

3. C. L. Moore, Method for Evaluating U.S. Crude Oil Resources and Projecting Crude Oil Availability, U.S. Department of the Interior, Office of Oil and Gas, Washington, D.C., May, 1962, p. 36.

4. Ibid., pp. 38 - 41.

5. Ibid., p. 99.

6. Ibid., pp. 47 - 48.

7. Ibid., p. 99.

8. Ibid., p. 73.

SECTION FIVE - ESTIMATES FROM WELLS DRILLED

Many estimates of the likely benefits from exploration programs based on previous successful wildcat wells and discoveries from these wells have likely been made. Probably most of these investigations have not been published because they were based partly on a company's own experience and knowledge gained from their exploration activities.

Such studies are not strictly concerned with the overall rate of utilization within any political area. They are usually made as an aid in deciding among exploration programs which promise an adequate return on the capital invested.

Furthermore, the use of base data on discoveries (which may not yet have been fully appreciated to their ultimate value) from other regions may lead to serious errors in long term projections of the ultimate potential of a fairly new basin. Therefore, adjustments usually have to be devised to estimate the long term economics of production in the area.

Arps and Roberts postulated that in any sedimentary basin, the estimated number of additional pools, $\Delta\Gamma$ (of size between the limits Φ and $(\Phi + \Delta\Phi)$ in that basin was jointly proportional to the number of additional wildcats drilled, $\Delta\Psi$, the number of undiscovered pools of that size remaining, $\Gamma_0 - \Gamma$, and the average areal size of such pools, Φ, or

$$\frac{\Delta\Gamma}{\Delta\Psi} = c\,\Phi\,(\Gamma_0 - \Gamma) \qquad (3.5.1)$$

By putting Equation 3.5.1 in differential equation form, and integrating, Arps and Roberts obtained

$$\Gamma = \Gamma_0 \quad (1 - e^{-c\Phi\Psi}), \qquad (3.5.2)$$

where

Φ = average areal size of the pools being considered,

Γ = the number of such pools discovered after a given number of wildcats, Ψ, had been drilled in the basin,

Γ_o = the original number of such pools before drilling had started,

Ψ = the number of wildcats drilled in the basin, and

c = a constant, depending upon exploration technique.

It should be noted that Equations 3.5.1 and 3.5.2 are not dependent upon time.

Arps and Roberts classified pools into fifteen groups according to their estimated ultimate recovery. The limits for each group were chosen so that each interval was twice as large as the interval for the preceding group. The smallest group had an estimated average ultimate recovery of 1,700 barrels, and the largest group had an estimated average ultimate recovery of 50,750,000 barrels of crude oil.

The average productive area of pools was calculated for each group using the relationship

$$\mu = 530 \ (\Phi)^{1.275}, \qquad (3.5.3)$$

where μ = estimated average ultimate recovery of pools,

and Φ = average areal extent of pools in acres.

A (smoothed) histogram of the frequency of the 338 pools as classified into fifteen groups formed the basis for estimating the effect of drilling additional wildcats; Equations 3.5.2 was used to estimate the incremental number of pools (for each of the fifteen groups) which would likely be found by drilling a given number of wildcat wells.

Arps and Roberts also analyzed the data on the average number of dry holes, N, drilled during development of fields to delineate pools of average areal size Φ acres, and found that

$$N = 0.76 \ \Phi^{0.345}, \qquad (3.5.4)$$

closely approximated the relationship.[1]

Kaufman provided a probabilistic adaptation to the deterministic model proposed by Roberts and Arps. Kaufman assumed that the volumetric size of pools within a field (or fields) would be lognormally distributed if all pools were found. The bivariate lognormal distribution was assumed to describe the areal extent in acres, and average net feet of pay of all pools which might eventually be found in the field.

If the above assumptions were true, then when a number of pools of various sizes had been discovered in a field, an estimate of the number and sizes of pools remaining to be found could be made. Kaufman postulated that by studying successive frequency distributions of pool sizes for several stages during the development of a field, some idea of what might be expected if more wells were drilled might be obtained. Also, if the true size of the overall field was derived by geological estimates and by applying subjective judgements, the eventual distribution (which would be asymptotic to the lognormal) might be estimated.[3]

Kaufman planned to study some of the problems of projection in cooperation with Paul Bradley of the University of British Columbia.[3] Mr. Crabbé of the National Energy Board is also studying this problem.[4]

Because of the lack of any other means of estimation, several studies have employed the results obtained in fairly well explored areas of the world, to assess potential resources of oil elsewhere. The authors of these studies have almost invariably warned of the gross errors which may be inherent in such estimates; the following quotation from a report by T. A. Hendricks clearly states some of the problems involved:

"Geological science has been very successful in classifying parts of the earth's crust as favorable, unfavorable, or impossible for the occurrence of oil and gas, but the exact location and extent of the occurrences can be determined only by drilling. By the same token, only extensive drilling and production experience can provide the statistical sample needed to attempt quantitative estimates of the producible quantities in the unexplored favorable parts of the earth's crust. Moreover, much of the purely geologic information needed for intelligent appraisal of unexplored rock can be provided only by drilling.

"At present, only a tiny part of the geologically favorable rock in the earth's crust has been explored by drilling. The amount of exploratory footage drilled in the United States dwarfs the total footage in the rest of the world, but even in the United States there is "room" in the geologically favorable rocks for more than six times as much exploratory drilling as has been done so far. Until a far larger "sample" has been drilled, estimates of undiscovered quantities will be subject to major revision".[5]

However, Hendricks pointed out that some estimate of the ultimate amount of crude oil in place could be obtained if the following factors were known, or could be estimated:

1. The quantity of oil in place found by exploration to date.
2. The fraction of the potentially productive rocks that have been explored to date.
3. The estimated incidence of oil in the unexplored part of favorable rock, as determined by comparison with the explored part.[6]

The actual amount of crude oil in the ground would be related to a number of factors. With 328 billion barrels of oil in place having been reported as of 1960, Hendricks assumed that there must be at least 400 billion barrels in known producing fields in the United States, which would eventually be reported through extensions and revisions.[7]

Hendricks then grouped oil producing states into three categories, according to their productiveness. He assumed the incidence of oil in the unexplored areas of each category to be:

1. equal to the average of all explored areas, (i.e. the average for all states) for the most productive states.
2. zero for the least productive states.
3. the average of 1. and 2. for all other states.

On this basis, Hendricks arrived at an estimate of 1,600 billion barrels of crude oil in the ground in the United States (excluding Alaska).[8]

Assuming that one third of the (five sixths) remaining unexplored area would be drilled before exploration became uneconomical and ceased, and further assuming that the incidence of oil in the areas to be drilled would be 75% of that in the explored areas, Hendricks finally arrived at an estimate of 1,000 billion barrels for the ultimate resource of oil in place. [9]

With the above figures for the United States as a basis, Hendricks went on to estimate the world oil in place at 10,000 billion barrels, of which 6,200 billion barrels would ultimately be recoverable. This estimate was based on a number of assumptions about the relative productivity of various areas of the world in comparison to the three major categories (according to productivity) established by Hendricks for the United States. [10]

Weeks contended that "petroleum resource estimates, ... can only be determined by correlating the facts of oil occurrence with geology, as revealed by the worldwide exploration experience of the industry". He points out that the incidence of oil varies from an average of less than one thousand, to several million barrels of oil per cubic mile of sediments.

According to Weeks, the only basis for resource estimates is that of "exploration experience on the widest possible basis, scientifically analyzed, (and) supplemented by realistic research". He stated that "extrapolations of production curves", and "indiscriminate comparison of areas or volumes of sediments", may produce grossly misleading figures. [11]

Weeks suggests that the most valid indication of oil in a particular sedimentary formation can be obtained from a knowledge of its age and structural history. Seepages, other surface evidence of oil, geophysical data, groundwater movement, and results from mapping techniques may give some further clues as to occurrence, and yet, to quote an axiom of the industry, "oil is where you find it".

Weeks also suggests that some scientific means must be developed to correlate all the factors which relate to the occurrence of oil. Most of Weeks' studies have not been concerned with trends in percent recovery, or on the different political and economic factors, which (while they have no bearing on oil incidence), certainly must relate to the dynamics of utilization. [12]

With the experience of these studies as an aid, the task which remains is to develop a model of the ultimate utilization of oil which would tend to embody some of the political, economic, technological and geological factors, (assuming these factors can be expressed implicitly). This has been attempted in the next Chapters.

REFERENCES:

1. J. J. Arps and J. G. Roberts, Economics of Drilling for Cretaceous Oil on East Flank of Denver - Julesburg Basin, Bulletin of the American Association of Petroleum Geologists, Volume 42, Number 11, November, 1958, pp. 2549 - 2566.

2. Gordon M. Kaufman, Statistical Analysis of the Size Distribution of Oil and Gas Fields, (unpublished paper), Massachusetts Institute of Technology, Cambridge, Mass., 1969, pp. 109 - 124.

3. Gordon M. Kaufman, Personal letter, April 30, 1969.

4. P. J. Crabbé, National Energy Board, Ottawa, interview, March, 1969.

5. T. A. Hendricks, Resources of Oil, Gas, and Natural-Gas Liquids in the United States and the World, Geological Survey Circular 522, United States Department of the Interior, Washington, 1966, p. 2.

6. Ibid., p. 3.

7. Ibid., p. 7.

8. Ibid., pp. 8, 15.

9. Ibid., pp. 9 - 10.

10. Ibid., pp. 13, 17.

11. Lewis G. Weeks, Worldwide Review of Petroleum Exploration, Excerpt, Sixth World Petroleum Congress Proceedings, Frankfurt, June, 1963, p. 261.

12. Ibid., pp. 261 - 267.

SECTION SIX - MULTIVARIATE MODELS OF DEMAND

If one accepts the assumption that the ultimate resource of oil (R) is finite, there are few (if any) exogenous variables which can be linearly related to the domestic supply of oil. Similarly, any transformation of such variates in order to achieve a one-to-one relationship to y (or $\hat{y}$) would require some Herculean assumptions. For this reason, multivariate models are usually used for projections of energy demand rather than for estimating crude oil supply.

So far as each firm is concerned, such studies are certainly necessary in order to project demand for crude oil and petroleum products. Usually, the time span for these projections is restricted to a period of less than thirty or forty years. For projections of demand for products (such as gasoline) a much shorter time base would likely be used.

Probably a large number of (short term) projections of crude oil and product demand have been made. It is instructive to describe the technique used in multivariate analysis, as illustrated by a recent study of the United States Department of the Interior. Some notion of the relationship between variables can be obtained, even though this writer feels that domestic crude oil production could be treated as an independent variable rather than a dependent variable--at least for a long term projection in which a substantial portion of the ultimate resource R would be produced.

In the Department of the Interior study variables were divided into two groups, "the independent (exogenous) variables, and the dependent (endogenous) variables". Some variables were transformed logarithmically or exponentially.

The independent variables were believed to be:

X 1= Economic activity

X 2= Population

X 3= Industrial production

X 4= Real costs and prices of energy resources

X 5= Domestic supply

X 6= Foreign trade

X 7= Environmental restrictions

X 8= Evolutionary technology

X 9= Revolutionary technology

X10= Regional factors

X11= Energy policy

X12= Political considerations and trade-offs

X13= Other variables accepted as given

The basic model for total energy consumption was $Y1 = f(X1, X2, X3, \ldots, X13)$, where Y1 was the dependent variable representing total energy demand.

The major components and subcomponents of the dependent variable for total energy were:

Y 1= Total energy consumption,

Y 2= Demand for bituminous coal and lignite,

Y 3= Demand for anthracite,

Y 4= Production of crude petroleum,

Y 5= Demand for petroleum and natural gas liquids,

Y 6= Production of natural gas, wet, including natural gas liquids,

Y 7= Demand for natural gas, dry

Y 8= Demand for hydropower, and

Y 9= Demand for nuclear power,

where $Y1 = Y2 + Y3 + Y5 + Y7 + Y8 + Y9$.

The major energy markets or consuming sectors were:

Z 1= Household and commercial sector demand

Z 2= Industrial sector demand,

Z 3= Transportation sector demand,

Z 4= Electric utilities sector demand, and

Z 5= Miscellaneous and unaccounted for,

where $Y1 = Z1 + Z2 + Z3 + Z4 + Z5$.

The major forms of energy consumption were:

N 1= Direct fuel uses,

N 2= Utility electricity uses, and

N 3= Raw material nonfuel and nonpower uses,

where $Y1 = N1 + N2 + N3$.

A number of composite cases were considered, including an analysis of demand for energy resources. In carrying out the projection to the year 1980 for crude oil production, assumptions were made regarding the percentage change anticipated in each of the independent variables for the period from 1965 to 1980; these assumptions were based on the results from the 1947 to 1965 data, and on subjective judgements.[1]

REFERENCE:

1. W.E. Morrison and C.L. Readling;, An Energy Model for the United States, Featuring Energy Balances for the Years 1947 to 1965 and Projections and Forecasts to the Years 1980 and 2000, Bureau of Mines Information Circular 8384, United States Department of the Interior, 1968, pp. 12-13, pp. 18-22, pp. 29-30, and p. 123.

CHAPTER IV

SOME MODELS OF THE DISCOVERY AND UTILIZATION PROCESS

SECTION ONE - PROBLEMS OF MODEL BUILDING

The application of the scientific method to the solution of any problem initially involves formulation of the problem. The mere statement of a problem necessarily implies the kind of solution desired - otherwise the problem could not have been formulated in the first place.

Problems which are never defined do not form the basis for applied scientific endeavors. On the other hand, the moment a problem is defined, the definition essentially comprises the first notion of the model which can be used to obtain an optimal solution to the problem.

As any scientist (and practical person) knows, it is only common sense that his initial concept about a situation should be tested against some facts. This step will either prove the original notion to be true in both diagnosis and fact, or will lead to a more refined statement of the problem.

This process can be repeated (perhaps over and over) until a clear theoretical understanding is obtained of both what the problem, and the solution, might be. The sort of intellectual activity involved, in searching (and researching) for an optimal solution, can be described as a three-step process.

A. Design the Model: In the beginning, the statement of the problem provides the initial framework for the design of the model, because the principles which apply to a solution are implicit in the formulation of the problem.

B. Test the Model: The facts which apply here can sometimes be drawn from experience. However, in most scientific studies, observations and actual data are usually necessary to describe with any accuracy what happens, where it happens, and when it happens.

C. Refine the Model: The facts gathered in the previous step can form the basis of how one's concept may be reformulated. Thus, an increasing awareness of why events occur in a certain way may lead to a redefinition of the model, and a return to Step A above.

This process may be repeated a number of times. Someone (identity unknown) once remarked that this iterative procedure could be called the "Las Vegas Method", in which improvements are made to the model until the benefits derived are less than the actual cost of making further improvements.

In mathematical analysis, the computer can often carry out much of the optimalization* process. The final polarization of the problem into an analytically consistent, and succinctly descriptive method provides the basis for the next stage - that of generalizing the methodology for the use of others.

This experiment of actually trying the solution, is sometimes called the "acid test", paraphrased by "Yes, but will it fly?". If the experiment is a success, then perhaps others can obtain similar results. Otherwise, it is a case of "back to the drawing board", to design a model which will predict satisfactorily.

In the pure sciences, much theoretical framework can be built up without necessarily experimenting to test the results. Here, the models developed may provide a conceptual basis to describe a variety of events. However, in the applied sciences, where concrete objectives (or goals) are specified, the solution to the original problem can only be attained by application of methodology in the real world.[1]

No doubt, the growing sophistication of technology plays a role in model development and utilization of models. Changes which occur in technology may bring about a need of refinement (or even realignment) of old theories. Conversely, new facts may find an unexpected home in theories developed many years ago.

Economic changes, and changes in human values, also bring about reemphasis (and even restructuring) in the models used. When these changes occur, the process of model building is set in motion again. Here, revisions in the model may be necessitated by the overriding influence of some variable previously unrecognized, or perhaps discounted in importance.

* Optimalization: Defined herein to be a systematic search of feasible solutions to find one which maximizes (or minimizes) an objective.

The steps from the definition to the solution of problems, by means of the scientific method have been illustrated in Figure 4.1.1. Application of common sense to everyday problems is very similar in nature to the use of the scientific method - excepting in the rigor of analysis required for model building.

Occasionally, the task of searching for optimal solutions has been tackled using the black box concept. Here, given a clear definition of objective, the computer carries out the tedious job of finding the optimal solution. By means of feed-back mechanisms, the computer can even be programmed to take over the decision making process. While this is a necessary and valuable expedient in some situations, sooner or later, what goes on inside the black box becomes of interest.[2,3]

For the present study, the choice in model building was to favour traditional causal models. More complicated systems could be suggested in the light of any fundamental relationships between variables which became established.

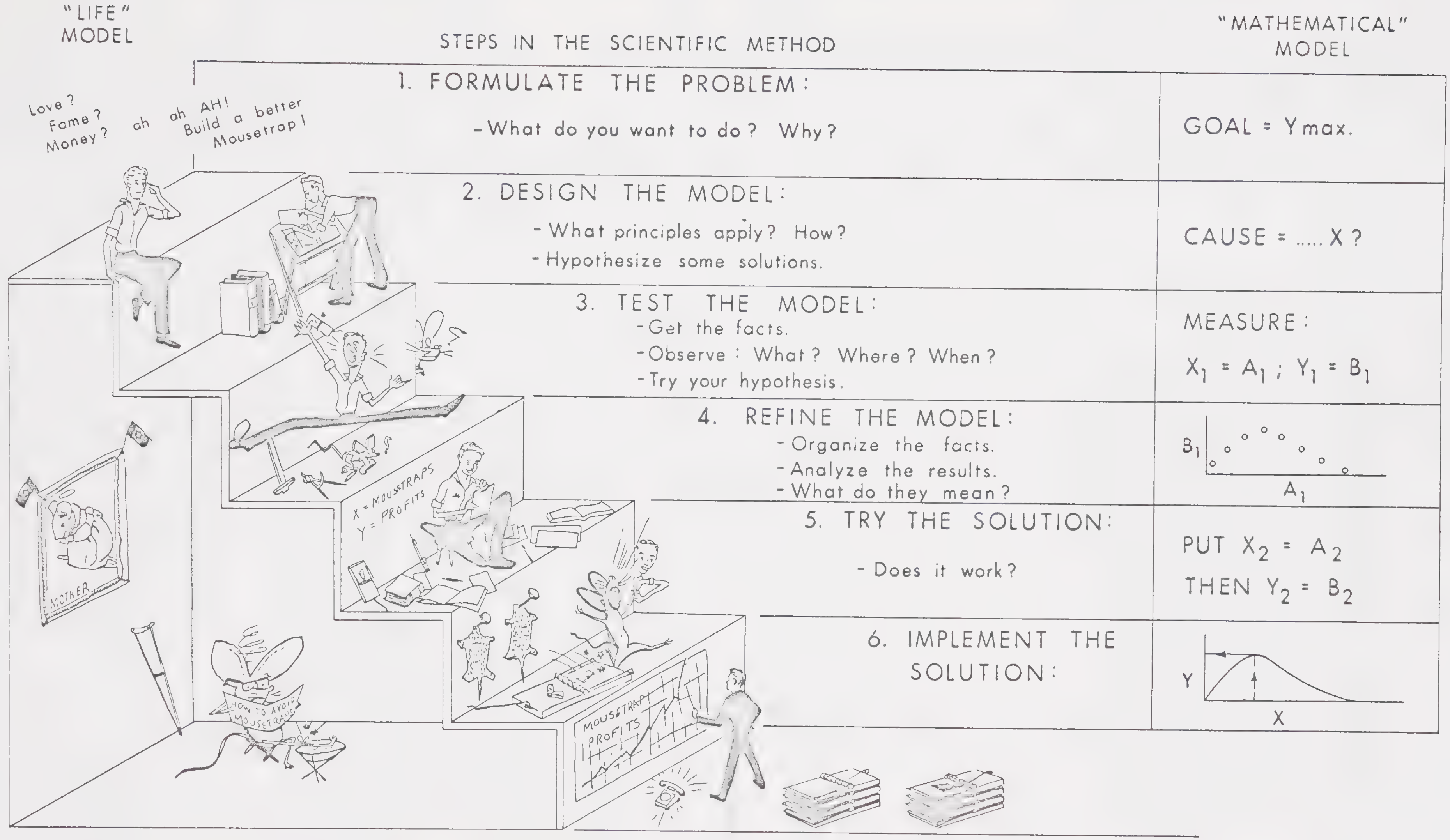
In a study such as the present one, with some benefit of hindsight from initial results obtained, it would always be possible to repropose some of the models in a more elegant fashion. However, the purpose of the present study was merely to determine the contribution of some variables believed to be significant. Other studies would be necessary to more fully define the relationship of all the predictive variables to the production of oil.

While there was a danger of failing to adequately depict the discovery and utilization of oil because an oversimplified model had been used, a first approximation might nevertheless be of some value.[4]

After some preliminary sketching of various functional relationships between the variables y, q, w, $\dot{y}$ and t, the task in this study became one of integrating these variables into a model which would serve as a comprehensive basis for estimating the recoverable resource of oil. Data for the United States were chosen for use in the models because of the amount of information available, and because other studies could be used in checking the order of magnitude of the results.

FIGURE 4. 1. 1

MODEL BUILDING AND THE SCIENTIFIC METHOD



Some other areas in which previous studies had been made included:[5,6]

1) Venezuela, and

2) Alberta.

It was decided that any models developed would be applied to only the United States in the present study. As described in Chapter Two, the oil industry is international in nature; more meaningful results might be obtained using the statistics for the following regions in the models:

1) World,
2) Western Hemisphere,
3) North America,
4) South America, and
5) Eastern Hemisphere.

The regional models should ideally be additive functions; ultimate resource projections for regions should add up to the projection for all the regions.

In this study, the models favoured were primarily deterministic in nature, because it was assumed that the recoverable resource of oil was

1) some finite value, and
2) determinate, in the sense that cumulative discoveries and cumulative production would (eventually) asymptotically converge on this value.

Thus, it was assumed that a deterministic model would provide a basis for the discovery and utilization process. Variation of this model from the data would therefore be described stochastically - just as other purely random events which can be described stochastically, such as the fall of the dice or the spin of the roulette wheel. The deterministic model implicitly assumes that for a projection of the process, all causal factors or conditions would exert a continuing uniform effect on the process.

The adequacy of a deterministic model should become apparent from an analysis of the randomness of variations. If the variations could not be suitably explained by conventional statistical techniques, then the basic analytical model might require further modification.

The method of least squares was assumed to offer the most advantages in testing the goodness of fit of the models.[7]

The least squares principle

1) leads to simpler statistical calculations and results,
2) does not involve any other assumption about the distribution of the variations,
3) proved to be essentially practical to use in computer algorithms even with non-linear relationships, and for this study
4) in most models tended to emphasize the more recent (and probably more reliable data available) because of the general increase in the magnitude of variables with time.

Computer algorithms which sought out the parameters in non-linear equations for a least squares fit were developed. Care was exercised in relating variables, so as to incorporate the notion of cause and effect in the model.

The regression between variables does not establish that a causal relation exists. Indeed, no amount of observation alone could assure that cause and effect are at work. But without some kind of model of the situation, the relevance of variables cannot be tested.

One hazard in model building is a tendency to rely on empirical or descriptive models. Descriptive models (born out of convenience) have been used frequently to characterize short-term dynamics. However, for long-term projections, it is far better to establish causal analytic models, which match any a priori knowledge of the situation and accommodate inferences about the nature of the process.[8]

At this point in the study, it was useful to adopt the following notation for the main variables used in describing the process of discovery and utilization of crude oil:

R = Ultimate recoverable resource

y = Cumulative production at the end of the year t,

$= y(t)$, if the year t is used as an index.

$\dot{y}$ = Production rate,

$= \dfrac{dy}{dt}$

$\frac{\Delta y}{\Delta t}$ = Annual production during the year t,

= $\Delta y(t)$, if the year t is used as an index,

= Δy for convenience, since $\Delta t = 1$ year.

r = proved reserves at the end of year t,

= r (t), if the year t is used as an index.

q = Cumulative discoveries at the end of year t,

= q(t), if the year t is used as an index,

= y + r

$\dot{q}$ = Discovery rate,

$= \frac{dq}{dt}$

$\frac{\Delta q}{\Delta t}$ = Annual increase in cumulative discoveries,

= Δq (t), if the year t is used as an index,

= Δq for convenience, since $\Delta t = 1$ year.

u = Proved discoveries classified by year of discovery, as defined in Equation 3.3.3 (= u(t), with t as an index).

w = Cumulative discoveries of oil in place at the end of year t, as defined in Equation 3.4.1,

= w (t), if the year t is used as an index.

t = An arbitrary time base, in years.

N = The year A.D. in which production was first started (or recorded).

i, j, or k designated the year A.D.

a, b, c, d, ... were parameters in the equations used.

Wherever possible equations have been written with a, b, c, d, ... >o.

In this study, a parameter was defined as an arbitrary constant or coefficient in a mathematical expression, which distinguished various specific cases. For example, in q = m y + c, m and c were the parameters, and they could be estimated by means of the regression of q on y.

A structural parameter was defined as a parameter in an equation which met certain boundary conditions. The structural parameters for such models could be estimated either by regression, or by some least squares technique, provided the boundary conditions were met. For example, in the equation $\dot{y} = my$, with $y = o$ for $\dot{y} = o$, the regression of $\dot{y}$ on y would be through the origin. It was assumed that the structural parameter R could be estimated from any model viable in long term projections.

Some assumptions proposed for analytical model were as follows:

1) y is a monotonically increasing function of time
2) R is some finite value.
3) $0 \leq y \leq R$.
4) $0 \leq q \leq R$.
5) $\dot{y} \geq 0$ for some finite time interval t, $0 \leq t \leq s$.
6) If $t = 0$, $y \doteq 0$.
7) If $t = s$, $y \doteq R$.
8) If $y = R$, $q = R$, and $\dot{y} = \dot{q} = 0$.
9) A single maximum value exists for $\dot{y}$.
10) A single maximum value exists for $\dot{q}$.
11) $y = \int_o^t \dot{y}\, d\, t$
12) The functional relationship between the variables y, q and t could be expressed in a reasonably simple form.

For the stochastic portion of the model, the following assumptions were made:

1) observed values of Y(i), Q(j), and W(k), were available for consecutive years,

$$i = N,\ N + 1,\ N + 2,\ \dots.\ L,$$
$$j = M,\ M + 1,\ M + 2,\ \dots.\ L,$$
$$k = P,\ P + 1,\ P + 2,\ \dots.\ L,$$

where Y(i) = cumulative production measured for every year up to year i,

$Q(j)$ = cumulative discoveries measured up to year j, and

$W(k)$ = cumulative discoveries of oil in place measured up to year k.

2) $y(t)$ and $\dot{y}(t)$ were the calculated values corresponding to the observed values $Y(i)$ and $\dot{Y}(i)$.

3) $q(t)$ and $\dot{q}(t)$ were the calculated values corresponding to the observed values $Q(j)$ and $\dot{Q}(j)$.

4) $w(t)$ and $\dot{w}(t)$ were the calculated values corresponding to the observed values $W(k)$ and $\dot{W}(k)$.

5) $Y(i) = \sum_{t=N}^{i} \dot{Y}(t)$, and $W(k) = \sum_{t=P}^{k} \dot{W}(t)$.

6) $Q = Y + r$.

7) The deviations (δ , Δ, and D) of the observed values from the calculated values were normally distributed,

where
$$\delta = \dot{y} - \dot{Y},$$
$$\Delta = y - Y, \quad \text{and}$$
$$D = q - Q.$$

Note that since $q = y + r$ and $Q = Y + r$, then $D = \Delta$.

One might expect that the relative deviations of Y and Q from their calculated values (y and q) to be subject to some kind of damping if data were available for y and q as they approached R. For this reason, the magnitude of the deviations (Δ and D) were assumed to provide a reasonable basis upon which to derive the parameters in a least squares computer algorithm.[9]

The deviations in annual production (δ), were similarly assumed to be random, and independent. Now, the rate equation for production ($\dot{y}$) is a logical consequence of the equation expressing cumulative production (y) as a function of time. Therefore, the projection of y and $\dot{y}$ to their ultimate values should lead to essentially the same result for R, or a rejection of the assumption that a particular equation describes the rate of utilization of oil.

Any causal relationships which have been proposed have been expressed by designating the dependent and the independent variables; independent variables usually appear on the right hand side of an equation. In the actual regression equation the error term (residual) would be added to the right hand side of the equation. It was assumed that the relative merits of each model could be estimated by various tests on the distribution of the residuals. These tests are described in the next Chapter.

REFERENCES:

1. Russel L. Ackoff, Scientific Method Optimizing Applied Research Decisions, "The Nature of Science and Methodology", John Wiley & Sons, Inc., 1962, pp. 1 - 28.

2. M. G. Kendall, Model Building and Its Problems, Mathematical Model Building in Economics and Industry, Conference, London, Hafner Publishing Company, New York, 1968, pp. 3 - 5.

3. Herbert A. Simon, The Sciences of the Artificial, "The Architecture of Complexity", The M.I.T. Press, Cambridge, Mass., 1969, pp. 84 - 118.

4. H. O. Wold, Model Building and Scientific Method A Graphic Introduction, Mathematical Model Building in Economics and Industry, Conference, London, Hafner Publishing Company, New York, 1968, p. 144.

5. Anibal R. Martinez, "Our Gift, Our Oil", N. V. Drukkerij, D. Reidel-Dordrecht, Netherlands, 1966, pp. 64 - 78.

6. Jack Meneer, Potential Energy Resources of Alberta and Its Strategic Positions Within the Circular Mosaic, Paper No. 6907, (Preprint), 20th Annual Meeting of the Petroleum Society of C.I.M., Edmonton, May, 1969.

7. Harald Cramer, The Elements of Probability Theory and Some of Its Applications, John Wiley & Sons, New York, 1955. pp. 133, 236.

8. Kendall, Model Building..., pp. 4 - 5.

9. F. R. Oliver, Aspects of the Maximum Likelihood Estimation of the Logistic Growth Function, American Statistical Association Journal, September, 1966, pp. 697 - 705.

SECTION TWO - MODELS DEPENDENT UPON TIME
- CONTINUOUS TIME FUNCTIONS

Long term models of the rate of utilization of oil must be bounded in cumulative production (or cumulative discoveries) as t becomes large. Unbounded models impute an estimate for R which does not exist. Some restriction must therefore apply to the magnitude of y (or q) in any model used to estimate R when time is the independent variable.

The usual method for limiting y (or q) has been to choose an equation which is asymptotic to some finite value as $t \to \infty$. Alternatively, any expression for $\dot{y}$ (or $\dot{q}$) can be chosen, provided the general requirements described in the previous section are met.

The annual statistics for Δy would provide a reasonable approximation for $\dot{y}$ in such equations. This approximation was useful for the derivation of comparable estimates of R, using different forms of the same basic model. A model in differential equation form should result in roughly the same estimate of R as would be obtained from the integrated form.

In this Section, five analytical models which met most of the conditions outlined in Section One of this Chapter are described. With time as the independent variable, all these models were non linear. Because of problems which may be encountered in estimation of the parameters in the integrated equation, the models have also been expressed in differential equation form (and where possible, as a difference equation)

1. Exponential Growth: The hypothesis made for the case of exponential growth was that the rate of production is a constant proportion of cumulative production so that

$$\dot{y} = c\,y \qquad (4.2.1)$$

The general solution of this equation is

$$y = k\,e^{ct}, \qquad (4.2.2)$$

or $$\ln y = \ln k + c\,t, \qquad (4.2.3)$$

where c and k are positive constants.

Equation 4.3.3 is linear in $\ell n\, y$, and the constants c and $\ell n\, k$ can be estimated by means of linear regression. In exponential growth, an upper limit for y does not exist as $t \to \infty$. Therefore this function does not provide an estimate for R, and t must be restricted to a finite time horizon.

In exponential decline in growth, y is a decreasing function of time. With an initial value of y given by $y\,(o) > o$ for $t = o$, Equation 4.2.2 then becomes

$$y = y\,(o)\, e^{ct}, \qquad (4.2.4)$$

with $c < o$.

In this case the function is bounded, and as $t \to \infty$, $y \to o$.

In a model for decline in growth (such as proposed by Roberts and Arps for the success of wildcat wells, and described in the previous Chapter) a hypothesis can be made that the rate of change is a constant proportion of the amout remaining, so that

$$\dot{y} = c\,(A - y), \qquad (4.2.5)$$

where A was the original amount which existed when $t = o$. Integrating this expression with the boundary condition $y\,(o) = o$ for $t = o$, we have

$$y = A(1 - e^{ct}), \qquad (4.2.6)$$

which is of the same form as the function used by Roberts and Arps. In this case, as $t \to \infty$, $y \to A$ for $c < o$.

Here, linear regression cannot be employed in the manner described for estimating the constants in Equations 4.2.2 and 4.2.3. A least squares computer algorithm could be used to estimate the coefficients of the equation. Also, linear regression of the discrete time function given by Equation 4.3.1 would provide an estimate of the constants of the equation.

2. <u>The Gompertz Curve:</u> The Gompertz curve has an equation of the form

$$y = Rc^{d^t} \qquad (4.2.7)$$

where $o < c < 1$ and $o < d < 1$. As $t \to \infty$, $y \to R$. The value of y at $t = o$ is Rc, and as $t \to -\infty$, $y \to o$. In this double exponential equation the increments in $\ell n\, y$ (if t increases in equal increments) are proportional to $\ell n\, y$, as shown later in this Section.[1,2]

The curve can be modified so as to have the lower asymptote equal to some positive value, say λ, and

$$y = \lambda + \gamma e^{-e^{-(\alpha + \beta t)}} \qquad (4.2.8)$$

where $R = \lambda + \gamma$ would be the upper asymptote,

α = a parameter determining the time origin, and

β = the time unit (or rate) parameter.

This latter expression has been called the modified Gompertz curve.[3]

By differentiating the equation for the Gompertz curve we obtain

$$\dot{y} = k\, y\, d^t = k\, y\, e^{-at}, \qquad (4.2.9)$$

or $$\dot{y} = -a\, y\, \ln (y/R), \qquad (4.2.10)$$

with $k = (\ln c)(\ln d)$, and $a = -\ln d$. Since $d < 1$, $a > 0$. The above equations are bounded using reasonable limits for y and t. (e.g. $y \neq 0$ for $t = 0$).

From Equation 4.2.10, the second derivative of y is

$$\ddot{y} = -a\, \dot{y}\, [\ln(y/R) + 1] \qquad (4.2.11)$$

From this expression, it can be seen that the maximum value of $\dot{y}$ occurs when $y = R/e$. However, there seemed to be no a priori reason to assume (as one would in adopting the Gompertz equation) that annual production should reach a maximum value when 36.9% of the resource had been produced. The maximum value in $\dot{y}$ would probably be determined by other factors (such as discovery rate, cumulative discoveries, and percent recovery) in addition to cumulative production.

The graph of $\dot{y}$ (when plotted against t) has a long tail to the right.[4] In the United States (excluding Alaska), where production seems to be reaching a maximum, most of the oil in place remains to be recovered, and new discoveries are still being made.[5] Gains in secondary recovery might be subject to some law of diminishing returns, as percent recovery tended to its ultimate maximum value. Also, the rate of discovery of oil would likely be subject to an exponential type of decline, as implied by the Arps and Roberts model for wildcat wells which was described in the previous Chapter.

The initial growth in $\dot{y}$ is similar to exponential growth, because of the term for y in the right hand side of the equation. The graph of $\dot{y}$ as a function of t has an inflection point when y is relatively small; it can be shown that inflection points

occur for $y \doteq 0.073R$, and for $y \doteq 0.682R$. From inspection of actual growth of oil production in a number of states in the Unites States, and other political areas, this would describe what actually happens in the early phase of development.[6]

Some transformations can be made of the Gompertz equation, so as to enable estimation of the parameters.

We can obtain from Equation 4.2.7

$\ell n\, [y\,(t+1)/y\,(t)\,] = d^t\,(\ell n\, c)\,(d-1),$

and $\ell n\, \ell n\, [y\,(t+1)/y\,(t)\,] = \alpha - at,$ $\qquad$ (4.2.12)

where $a = -\ell n\, d,$

and $\alpha = \ell n\, [\,(\ell n\, c)(d-1)\,].$

Equation 4.2.12 is linear in $\ell n\, \ell n\, [\,y\,(t+1)/y\,(t)\,]$. All the constants (except R) can be estimated in this equation by means of linear regression.

Similarly, from Equation 4.2.7,

$\ell n\, |\ell n\, (y/R)\,| = \gamma - at,$ $\qquad$ (4.2.13)

with $a = -\ell n\, d$

and $\gamma = \ell n\, |\ell n\, c|$

Equation 4.2.13 is linear in $\ell n\, \ell n\, (y/R)$. However, three parameters must be simultaneously determined in order to arrive at an estimate for R.

In order to retain R in the equation, put

$x = y + 1,$

$B = R + 2,$

and assume that for $t = 0$, $y = 0$, but that $x = 1$.

The upper limit of $B = R + 2$ is an arbitrary choice, but it was chosen so that $\ell n\, |\ell n(x/B)|$ would be bounded as $y \rightarrow R$, and as $x \rightarrow (B-1)$.

The Gompertz equation with this boundary condition takes the form

$x = Be^{-(\ell n\, B)\, e^{-at}}$ $\qquad$ (4.2.14)

The linear difference equation then becomes

$\ell n\, \ell n\, [x(t+1)/x(t)\,] = -at, \ell n\, [(\ell n\, B)(1-e^{-a})\,]$ $\qquad$ (4.2.15)

from which an estimate of R can be obtained, if $t = o$ when $x = 1$.

In the study by C. L. Moore, use was made of the fact that production data were available in consecutive annual increments.[7]

Thus, from Equation 4.2.7,

$\ln y = \ln R + d^t \ln c,$

and

$$\frac{d}{dt}(\ln y) = (\ln c)(\ln d)d^t$$
$$= (\ln d)(\ln y - \ln R)$$
$$= (\ln d)(\ln y) - (\ln d)(\ln R)$$

$$\therefore \Delta(\ln y)/\Delta t \doteq \omega - a \ln y \qquad (4.2.16)$$

where $a = -\ln d,$

$\omega = a \ln R$

and $\Delta t = 1$ year.

Thus, the annual increase in $\ln y$ is a linear function of $\ln y$; with regression analysis, the method would serve to minimize the sum of squares of deviations of the data from those calculated for $\Delta(\ln y)/\Delta t$.

A difference equation can also be derived for the equation of the Gompertz curve. From Equation 4.2.7 we can write

$\ln y = \ln R + d^t \ln c$

Now let $z(t) = \ln y,$ (4.2.17)

$b = \ln c,$

and $m = \ln R.$

Therefore $z(t) = m + b\, d^t,$

and $z(t+1) = m + b\, d^{t+1}.$

This pair of equations yields a difference equation

$\therefore z(t+1) - z(t) = b\, d^t (d-1),$

which can be written as

$$z(t+1) = m(1-d) + d\, z(t) \qquad (4.2.18)$$

This is a first order linear difference equation with constant coefficients. Estimates of the coefficients can be obtained by the regression of $\ln[y(t+1)]$ on $\ln[y(t)]$. Some other linear difference equations are described in the next Section.

Because of the logarithmic transformations of y, one would not expect the magnitude of deviations of Equations 4.2.12, 4.2.13, 4.2.15, 4.2.16 and 4.2.18 to be normally distributed. A computer algorithm in which (untransformed) deviations from the calculated values of $\dot{y}$ (or y) were minimized would be more likely to have residual errors which were normally distributed.

In any case, the results from each method were of interest. To the extent that these equations all led to the same estimate for R, there would be no practical reason for favoring one method over the other.

3. <u>The Logistic Curve:</u> The logistic curve has an equation of the form

$$q = R/(1 + A\, b^t), \tag{4.2.19}$$

where $o < b < 1$, and $A > o$. As $t \to \infty$, $q \to R$. The value of q at $t = o$ is $R/(1+A)$, and as $t \to -\infty$, $q \to o$. The curve is also known as the Pearl-Reed curve, and has been used extensively to describe population growth.[8]

The curve can be modified so as to have the lower asymptote equal to some positive value, say λ, and

$$q = \lambda + \gamma/\left(1 + e^{-(\alpha + \beta t)}\right) \tag{4.2.20}$$

Here, $R = (\lambda + \gamma)$ would be the upper asymptote ,

α = a parameter determining the time origin, and

β = the time unit (or rate) parameter.

This latter expression has been called the modified logistic curve.[9]

Differentiating Equation 4.2.19, one can obtain

$$\dot{q} = k\, q^2\, b^t = k\, q^2\, e^{-at}, \tag{4.2.21}$$

where $k = a\,A/R$,

and $a = |\ln b|$.

Also, $$\dot{q} = c\, q\, (R - q), \tag{4.2.22}$$

with $c = k/A$,

since $R - q = q\, A\, b^t$.

From Equation 4.2.22, the second derivative of q is

$$\ddot{q} = c\, \dot{q}\, (R - 2q). \tag{4.2.23}$$

One can see from this expression that the maximum value of $\dot{q}$ occurs when $q = R/2$. Furthermore, the graph of $\dot{q}$ plotted against t is symmetrical, with inflection points at $q \doteq (R/2)(1 \pm 3^{-\frac{1}{2}})$. There was no apparent basis for such a relationship to exist between $\dot{q}$ and t. While one might expect an exponential type of growth (when the relative effect of the $(R - y)$ term was small) there seemed to be no reason to assume that decline in production was exactly imaged by growth in production.

In order to determine the parameters in Equation 4.2.19, Hubbert plotted $\log[(R/q) - 1]$ for the data, using various values of R. The best (visual) linear relationship so obtained was selected as giving the best estimate for R, as outlined in the previous Chapter.

A more convenient method of estimating the parameters has been described by Tintner. From Equation 4.2.19 we can write

$$1/q(t) = 1/R + (A/R)\, b^{t}$$

and $$1/q(t+1) = 1/R + (A/R)\, b^{t+1}$$

Thus, $$1/q(t+1) - 1/q(t) = (A/R)\, b^{t}(b-1),$$

and $$1/q(t+1) = b/q(t) + (1-b)/R \qquad (4.2.24)$$

If we let $z(t) = 1/q(t)$, then Equation 4.2.24 can be expressed as the linear first order difference equation

$$z(t+1) = b\, z(t) + (1-b)/R, \qquad (4.2.25)$$

which Tintner derived in order to improve on the statistical methods of estimation of the parameters in the logistic equation. The coefficients b and $((1-b)/R)$ can be estimated from ordinary least squares by the regression of $z(t+1)$ on $z(t)$. [10, 11]

With b and R found in this way, the third parameter A can be estimated by the method of Rhodes. From Equation 4.2.19

$$q = R/(1 + Ae^{-\beta t}) \qquad (4.2.26)$$

with $\beta = |\ln b|$. Rewriting Equation 4.2.26 in terms of the reciprocal of q as before we have

$$1/q = 1/R + (A/R)\, e^{-\beta t}$$

and $$R/q - 1 = A\, e^{-\beta t}$$

so that $$\ln[(R/q) - 1] = \ln A - \beta t \qquad (4.2.27)$$

If β and R are known, then A can be calculated. Equation 4.2.27 can be stated for successive values of $t = 1,2,3,\ldots N$; by adding these equations and solving for ℓn A we obtain

$$\ell n A = \beta(N+1)/2 + (1/N) \sum_{t=1}^{N} \ell n \left[(R/q) - 1\right] \qquad (4.2.28)$$

from which A can be calculated.[12]

Another means of estimating the parameters is offered by a computer algorithm in which a least sum of squares of the residuals was obtained using Equation 4.2.19 (or Equation 4.2.22 with $\Delta q = \dot{q}$). The algorithm could simply seek out those values of the parameters which led to the best least squares fit with the data. This procedure offers the feature of specifying that the magnitude of the untransformed deviations of actual data from the calculated values for q (or $\dot{q}$) should be minimized.[13, 14]

The logistic and the Gompertz curves have implied quite different values for R; neither curve has produced consistent results using data for q and for y. This is unfortunate, because both q and y should tend to the same estimate for R, if all the resource of oil is eventually produced.

4. <u>The Generalized Logistic Curve:</u>

A problem inherent in the foregoing equations of the type $\dot{y} = k y f(y,R)$ would be that the proportionality constant (k) and the ultimate resource (R) were the only parameters. No parameter determined the increase in growth in relation to the decline in growth of $\dot{y}$ or $\dot{q}$. Thus, if one assumes that either the logistic, or the Gompertz curve provides the rule by which the resource of oil is utilized, then one implicitly assumes a unique relationship between the increase and the decline in $\dot{y}$, and in $\dot{q}$.

In order to overcome this problem, another parameter can be introduced so as to balance the growth and decline rate. In Equation 4.2.9 we had

$$\dot{y} = k y e^{-at} \qquad (4.2.9)$$

for the Gompertz case.

Similarly, in Equation 4.2.21 we had

$$\dot{q} = k q^{2} e^{-at} \qquad (4.2.21)$$

for the logistic case,

Thus, $\dot{y} = k y^{\alpha} e^{-at}$, $\qquad (4.2.29)$

and $\dot{q} = k q^{\alpha} e^{-at}$, $\qquad (4.2.30)$

with $1 < \alpha < 2$, are rate equations intermediate to the rate equations for the Gompertz and the logistic curves. Here α is a parameter which could serve to describe the inherent growth and decline characteristics.

This three parameter equation becomes a four parameter equation when integrated, since

if $\quad dy/y^{\alpha} = k e^{-at} dt,$

then $\quad (y^{-\beta})/\beta = c + (k/a) e^{-at}, \qquad (4.2.31)$

with $\quad o < (\alpha - 1) = \beta < 1$ in the intermediate case.

$\quad \therefore y^{-\beta} = 1/\gamma + (A/\gamma) e^{-at}$

with $\quad 1/\gamma = \beta c$, and $A = k/ac$.

$\quad \therefore y^{\beta} = \gamma/(1 + A e^{-at})$

and $\quad y = R/(1 + A e^{-at})^{\theta} \qquad (4.2.32)$

with $\quad R = \gamma^{\theta}$, where $1 < \theta = 1/\beta$. If $\theta = 1$, we have the equation for the logistic curve. For the Gompertz case, if $\alpha \to 1$, $\theta \to \infty$, and the equation is therefore intermediate to the logistic and the Gompertz.

With some rearrangement of terms, Equation 4.2.29 can also be written in the form

$$\dot{y} = (a/\beta) y \left[1 - (y/R)^{\beta} \right] \qquad (4.2.33)$$

provided $(\beta + 1) = \alpha \neq 1$. If we assume that $\beta > o$ in Equation 4.2.33, the range of values for θ is $0 < \theta < \infty$. If $\theta < 0$, the generalized logistic equation becomes a generalized exponential equation.[15] It was assumed that the estimation of the parameters in such equations could be obtained by means of a least squares computer algorithm.

5. <u>The Beta Function:</u> A rate equation with several parameters might provide more flexibility in describing the growth and decline process. For example, employing the function for the Beta distribution, we could put

$$\dot{y} = k\, y^{m} (R - y)^{n} \tag{4.2.34}$$

with $o < m < 1$, and $o < n < 1$. The integral of Equation 4.2.34 can be expressed in terms of a Beta function. If the resource of oil is produced in the span of time $o < t < s$, then the proportionality constant k is determined by the relationship [16]

$$k\, s = R^{1 - m - n}\, B\,[\,(1 - m),\,(1 - n)\,] \tag{4.2.35}$$

The parameters R, k, m, and n might be estimated by means of a least squares computer algorithm - provided the algorithm converged with this number of coefficients to be estimated simultaneously. The value of m and n which provided the best least squares fit would be of interest, as well as the value obtained for R.

If a three parameter equation was desired, then m (or n) could be set equal to some arbitrary value. The relative decline in growth could then be estimated independently from the value assessed for n. However, if the maximum in $\dot{y}$ is to occur before one half of the resource was produced, then $m < n$.

If we put $m + n = c$, Equation 4.2.34 is a homogeneous function of degree c. Furthermore, if $c = 1$, the equation is similar to an economic production function (except that y is the only variable on the right hand side). [17]

Thus $\ln \dot{y} = \ln k + m \ln y + (1-m) \ln (R - y)$, (4.2.36)

with $k\,s = \Gamma(1 - m)\;\Gamma(m)$.

Finally, if an assumption is made regarding the value for the time span(s), then Equation 4.2.36 has only two parameters (m and R) to be estimated.

However, there should be enough parameters in the equation for $\dot{y}$ (or $\dot{q}$) so that the following could be independently estimated:

i) the proportionality constant, k,

ii) the ultimate resource, R, and

iii) the relation between increase and decline in growth. (Alternatively,

the functional relationship of the variables on the right hand side could specify the nature of the increase and decline in growth). If other variables (such as q or w) were used in the right hand side in addition to y, then extra parameters might be necessary to link these with y.

REFERENCES:

1. C. L. Moore, Projection of United States Petroleum Supply to 1980, Office of Oil and Gas, Department of the Interior, Washington, D.C., 1966, Annex, p. 5.

2. Glenn James and Robert C. James, Mathematics Dictionary, Multilingual Edition, D. Van Nostrand Company, Inc., Princeton, N.J., 1959, p. 176.

3. J. C. Ahuja, Statistical Problems in the Study of Growth, Unpublished Ph. D. Thesis, Department of Mathematics, University of British Columbia, Vancouver, B. C., 1963, p. 1.

4. Moore, p. 8.

5. See Appendix A.

6. DeGolyer and MacNaughton, Twentieth Century Petroleum Statistics, 1969, Annual, DeGolyer and MacNaughton, Dallas, Texas, 1969, pp. 3 - 25.

7. Moore, pp. 5 - 6.

8. James and James, p. 240.

9. Ahuja, p. 9.

10. G. Tintner, The Logistic Law of Economic Development, Journal Paper No. J - 3977, Iowa Agricultural and Home Economics Experimental Station, Ames, Iowa, 4 pages, Reprinted from the Journal Arthaniti, Volume IV, Number 1, 1960.

11. G. Tintner, Econometrics, John Wiley & Sons, Inc., New York, 1952, pp. 208 - 209.

12. Ibid., p. 210.

13. See Appendix "Computer Programs Used".

14. F. R. Oliver, Aspects of the Maximum Likelihood Estimation of the Logistic Growth Function, American Statistical Association Journal, September, 1966, p. 697.

15. Ahuja., p. 76.

16. E. S. Keeping, Introduction to Statistical Inference, D. Van Nostrand Company, Inc., Princeton, New Jersey, 1962, pp. 83 - 85, and pp. 389 - 390.

17. Tintner, pp. 52 - 54.

SECTION THREE - MODELS DEPENDENT UPON TIME - DISCRETE TIME FUNCTIONS

1. First Order Difference Equation: The hypothesis made for a first order difference equation is that production during any year is a linear function of the production for the previous year, so that

$$y(t) = a\,y(t-1) + b, \qquad (4.3.1)$$

or

$$y(t) - a\,y(t-1) = b \qquad (4.3.2)$$

This is a linear difference equation of order one and with constant coefficients. The solution to such equations can be found using the Laplace transformation, which leads to a staircase function to represent the discrete changes which occur in y (t) for each designated time period.[1] However, the solution to Equation 4.3.2 is usually found by dropping the requirement that y (t) take on only discrete values for continuous time, and allowing a continuous expression for y (t) in which $t = 0, 1, 2, \ldots$

If y^* is a particular solution to the complete equation

$$y(t) - a\,y(t-1) = b, \qquad (4.3.2)$$

and if $y^\dagger$ is the general solution to the reduced equation

$$y(t) - a\,y(t-1) = 0, \qquad (4.3.3)$$

then

$$y(t) = y^* + y^\dagger \qquad (4.3.4)$$

will be the general solution of the complete equation.[2]

For Equation 4.3.3, the general solution is

$$y^\dagger = \lambda a^t, \qquad (4.3.5)$$

with λ an arbitrary constant.

For Equation 4.3.2, a particular solution is

$$y^* = A, \tag{4.3.6}$$

where $A = b/(1 - a)$ is a constant.

According to Equation 4.3.4, the general solution of Equation 4.3.2 is given by

$$y(t) = A + \lambda a^t \tag{4.3.7}$$

Now if $y(t) = y(o)$ for $t = 0$, the general solution can be written as

$$y(t) = A(1 - e^{ct}) + y(o)\, e^{ct}, \tag{4.3.8}$$

where $c = \ln a$. Finally, if $y(o) = 0$ for $t = 0$, then

$$y(t) = A(1 - e^{ct}) \tag{4.3.9}$$

For $c < 0$, $y(t)$ would be asymptotic to A as $t \rightarrow \infty$.

Equation 4.3.9 is identical in form to Equation 4.2.6. Thus, the coefficients in these equations can be estimated by means of the regression of $y(t)$ on $y(t - 1)$ as in Equation 4.3.1.

Finally, the coefficient c in Equations 4.2.1, 4.2.2 and 4.2.3 may be related to the coefficient a in Equations 4.3.1 and 4.3.2. If $y(t) = ke^{ct}$, as in Equation 4.2.2, then in Equation 4.3.2

$$ke^{c(t+1)} - a\, ke^{ct} = b \tag{4.3.10}$$

Now, if $c = \ln a$, then $b = 0$. Conversely, if the hypothesis $\hat{b} = 0$ is true, then we can infer that the constant c as determined in Equation 4.2.3 is equal to the constant $\ln a$, where $\hat{a}$ would be determined from Equation 4.3.1. Since the correlation between successive measurements depends upon their proximity, the probability that the hypothesis $\hat{b} = 0$ is true is improved if $\Delta t \rightarrow 0$, or if $\Delta t = 1$ and the range of t

is large .[3]

2. Second Order Difference Equation: The hypothesis made for a second order difference equation is that production during any year is a function of the production of the two previous years, so that

$$y(t) = c - a\,y(t-1) - b\,y(t-2) \qquad (4.3.11)$$

or

$$y(t+2) + a\,y(t+1) + b\,y(t) = c \qquad (4.3.12)$$

where a, b and c are constants for a linear difference equation .[4]

The corresponding homogenous or reduced equation is

$$y(t+2) + a\,y(t+1) + b\,y(t) = 0 \qquad (4.3.13)$$

As in the case of a first order difference equation, we attempt to find the general solution $y^\dagger$ to the reduced equation and a particular solution y^* to the complete equation; the general solution to the complete equation will then be

$$y = y^* + y^\dagger, \qquad (4.3.14)$$

For a trial general solution, let

$$y(t+n) = m^n \qquad (4.3.15)$$

be substituted in Equation 4.3.12 in order to obtain the auxiliary equation

$$m^2 + am + b = 0 \qquad (4.3.16)$$

The roots of this equation are

$$m = (-a \pm \sqrt{a^2 - 4b})/2 \qquad (4.3.17)$$

The roots m_1 and m_2 are not equal to zero, or Equation 4.3.13 would not be of second order. Thus, m_1 and m_2 are

i) real and not equal, and

$$y^\dagger = k_1 m_1^{\,t} + k_2 m_2^{\,t}, \tag{4.3.18}$$

ii) real and equal, and

$$y^\dagger = (k_1 + k_2 t) m_1^{\,t}, \tag{4.3.19}$$

or iii) complex conjugates, and

$$y^\dagger = k_1 r^t \cos (t\,\theta + k_2) \tag{4.3.20}$$

where $m = r\,(\cos \theta \pm i \sin \theta)$ are the roots in polar form. In these three cases k_1 and k_2 are arbitrary constants.[5]

A particular solution to Equation 4.3.12 is

$$y^* = \text{a constant}, \tag{4.3.21}$$

so that $y^* + a\,y^* + b\,y^* = c$

$$\text{and} \quad y^* = c/(1 + a + b) \tag{4.3.22}$$

Thus the general solution to Equation 4.3.11 is

$$y = y^\dagger + c/(1 + a + b), \tag{4.3.23}$$

where $y^\dagger$ is defined by Equations 4.3.18, 4.3.19 or 4.3.20.

3. <u>Polynomial in Time:</u> The hypothesis made here is that production is dependent upon time, and can be estimated by means of the polynomial

$$y\,(t) = \beta_0 t^0 + \beta_1 t^1 + \beta_2 t^2 + \ldots + \beta_m t^m + \delta, \tag{4.3.24}$$

where $t = 0, 1, 2, \ldots n$ are the years for which observations of the dependent variable $y\,(t)$ were available, β_i = the unknown population regression coefficients, $i = 0, 1, 2, \ldots m$, $m < n - 1$, and δ = the error term in the equation.

Since the years $t = 0, 1, 2, \ldots n$ are not assumed to be independent,

t can be expressed in a power series as above, with each term t^i being considered as a new variable (say t_i) in the regression equation. Considered in this way, estimators of the regression coefficients ($\hat{\beta}_i$) can then be obtained by means of linear regression.[5,6]

If the independent variables are introduced in the order, $t, t^2, t^3, \ldots t^m$, then the reduction in the residual sum of squares caused by successively higher order powers may be estimated by means of analysis of variance. Similarly, from a polynomial of degree m in time, the contribution by each of the m terms can be estimated, and only those terms accounting for most of the variation included in the final equation. The details of this procedure have been described in a number of texts, and computer programs are available for the calculations.[7]

A polynomial of degree $n - 1$ would fit the n pairs of observations exactly, but there would be no degrees of freedom available to estimate the error mean square. Moreover, the inclusion of higher powers of t would not likely appreciably decrease the residual sum of squares. Therefore, some compromise is usually necessary in deciding on the order of polynomial to be used and the goodness of fit desired. In this study, terms up to the fourth power were included in the series. With $m = 4$, there would be some indication of the kind of relationship which might be anticipated between the successive coefficients βi, if $m > 4$ in an augmented power series.

A shortcoming of Equation 4.3.20 is that it provides no estimate for the parameter R, and is of use only for a bounded range of t. If $t \to \infty$, then $y(t) \to +\infty$, or $-\infty$, depending upon whether $\hat{\beta}_m$ is positive or negative.

However, Equation 4.3.20 can be used to arrive at some inferences regarding the residuals (i.e. the differences between the observed values, y(t), and the calculated

values, $y_c(t)$.) If we hypothesize a simple (parabolic) time pattern exists for the residuals, then we can put

$$\delta(t) = \beta_0 + \beta_1 t + \beta_2 t^2, \qquad (4.3.25)$$

where $\hat{\beta}_i$ will be the estimators of the population regression coefficients, as before. The residuals in Equation 4.3.25 are:

i) concave upward, if $\hat{\beta}_2 > 0$,

ii) concave downward, if $\hat{\beta}_2 < 0$, and

iii) not concave if $\hat{\beta}_2 = 0$.

The standard error of $\hat{\beta}_2$, and the reduction of the mean square for error by introducing the term $\beta_2 t^2$ in Equation 4.3.25 would indicate the degree of significance of such a trend .

REFERENCES:

1. Ruel V. Churchill, Modern Operational Mathematics in Engineering, Mcgraw-Hill Book Company, Inc., New York, 1944, pp. 23 - 27.

2. Samuel Goldberg, Introduction to Difference Equations, John Wiley & Sons, Inc., New York, 1961, p. 124.

3. J. C. Ahuja, Statistical Problems in the Study of Growth, unpublished Ph. D. Thesis, Department of Mathematics, University of British Columbia, Vancouver, B. C., 1963, p. 8.

4. Goldberg, p. 128.

5. Ibid., pp. 134 - 141.

6. E. S. Keeping, An Introduction to Statistical Inference, D. VanNostrand Company, Inc., Princeton, New Jersey, 1962, pp. 335 - 337.

7. K. W. Smillie, An Introduction to Regression and Correlation, The Ryerson Press, Toronto, 1966, pp. 78 - 82.

8. K. W. Smillie, STATPACK 2: An APL Statistical Package, Second Edition, Publication No. 17, Department of Computing Science, University of Alberta, Edmonton, Alberta, 1969, pp. 28, 33.

SECTION FOUR – RELATIONSHIPS OF PRODUCTION TO OTHER VARIABLES – ANNUAL PRODUCTION

To learn something of the relationship of annual production in the United States to other variables, a model which was proposed was

$$\Delta y(t) = \beta_0 + \beta_1 D(t) + \beta_2 G(t) + \beta_3 H(t) + \beta_4 N(t) + \beta_5 r(t), \quad (4.4.1)$$

where $\Delta y(t)$ = crude oil production in the United States during year t,

β_i = the unknown population regression coefficients, i=1,2,3,4,5,

$D(t)$ = consumption of all refined products in the United States during year t,

$G(t)$ = consumption of gasoline in the United States during year t,

$H(t)$ = crude oil production in the Western Hemisphere during year t,

$N(t)$ = crude oil production in North America during year t, and

$r(t)$ = proved reserve estimates for the United States at the end of year t.

Equation 4.4.1 is a multivariate equation in which the estimators of the coefficients $\hat{\beta}_i$ can be obtained by means of linear multiple regression. It was assumed that each variate did not require any transformation to reduce the equation to linear form – even for long term projections.

Because crude oil production in the United States has constituted a large portion of domestic production, and of production in the Western Hemisphere, a modification of Equation 4.4.1 was also proposed, as follows:

$$\Delta y(t) = \beta_0 + \beta_1 [D(t) - \Delta y(t)] + \beta_2 [H(t) - \Delta y(t)] + \beta_3 [N(t) - \Delta y(t)] + \beta_4 r(t) \quad (4.4.2)$$

This latter expression thus excludes production in the United States from the figures for production in the Western Hemisphere, and in North America. The term

$[D(t) - \Delta y(t)]$ is the difference between domestic demand for all refined oils and domestic production of crude oil in the United States.

No time lags have been used in the above models. A time lag might improve the relationship between production to proved reserves; the last term in the equations would then be $r(t - \tau)$, where τ is a time lag of say three to five years. Proved reserves are almost invariably produced in less than fifteen years - the average time depending upon the political region.[1,2,3]

REFERENCES:

1. M. King Hubbert, Degree of Advancement of Petroleum Exploration in United States, The American Association of Petroleum Geologists Bulletin, V. 51. No. 11, November, 1967, p. 2210.

2. Governors' Special Study Committee, A Study of Conservation of Oil and Gas In the United States, 1964, Interstate Oil Compact Commission, Oklahoma City, Oklahoma, 1964, p. 201.

3. World Oil, August 15th Issue, 1967, p. 40. (Also see August 15th Issue for other years.)

SECTION FIVE - RELATIONSHIPS OF PRODUCTION TO OTHER VARIABLES - CUMULATIVE PRODUCTION

In addition to the annual estimates of proved reserves, data were also available on the estimated amount of crude oil in place in the United States according to discovery year. Cumulative discoveries, and cumulative discoveries of oil in place are linked by the equation

$$p.r./100 = q/w, \tag{3.4.6}$$

according to the definition for percent recovery used by C. L. Moore.

Therefore, in order to use data on oil in place in a model which would provide an estimate for cumulative discoveries (which might in turn be used to estimate production) an expression for the trend in percent recovery in terms of t (or y) would have to be proposed. Percent recovery has usually been conceived as a time dependent function, as indicated in Chapter III. It has been mainly new technology which has brought about new techniques for secondary recovery, and subsequent increases in primary recovery.

In this study, a rather simple relationship was proposed to describe percent recovery, as follows:

$$p.r. = m\,y + b \tag{4.5.1}$$

Here, percent recovery is solely dependent upon cumulative production. It was implicitly assumed that as cumulative production increased, new techniques became not only possible, but necessary because of the diminishing amount of oil to be discovered. The equation had an additional appeal of expressing what might happen in the final stages of production - when there would be relatively few new discoveries of oil in place, and percent recovery would increase only as existing deposits were proved up and produced. An upper limit for R can be estimated from Equation 4.5.1, because the ultimate percent recovery must be less than 100%.

A model which was proposed to relate cumulative production directly to cumulative discoveries was

$$\frac{dy}{y} = c\,\frac{dq}{q} \tag{4.5.2}$$

This was an analytically simple function, which might serve to describe the build-up and final depletion of reserves. It satisfies the boundary conditions which were proposed in Section One of this Chapter, and expresses the concept that at the end of the decline process, each barrel of oil would be produced as soon as it was found.

If $\ln y = c \ln q + \ln b$

then $y = b\,q^c$,

$R = b\,R^c$,

$b = R^{1-c}$

$\therefore\ y = R^{1-c}\,q^c$

$= R^{1-c}\,(y+r)^c$

$\therefore\ y^{1/c} = R^{\frac{(1-c)}{c}}\,(y+r)$,

and $y + r = \dfrac{y^{1/c}}{R^{\frac{1-c}{c}}}$

When the Equation 4.5.2 is integrated, we obtain

$$\ln y = c \ln q + \ln b, \tag{4.5.3}$$

or $$y = b\,q^c, \tag{4.5.4}$$

with $b = R^{1-c}$, since if $y = q$, $r = o$, and $y = R$.

Equation 4.5.4 can also be stated as

$$(y/R) = (q/R)^c \tag{4.5.5}$$

with $c > 1$, since by definition $q = y + r$, and because proved reserve estimates are not negative. The hypothesis underlying Equations 4.5.2 to 4.5.5 is that incremental production (relative to cumulative production) is proportional to (but greater than) incremental discoveries (relative to cumulative discoveries).

Similar expressions were proposed for the relationships between cumulative production, cumulative discoveries of oil in place, and cumulative well footage drilled, as follows:

$$\ell n\, y = a\, \ell n\, w + k_1 \tag{4.5.6}$$

and

$$\ell n\, y = \alpha\, \ell n\, f + k_2 \tag{4.5.7}$$

where w (t) = cumulative discoveries of oil in place at the end of year t,

f (t) = cumulative well footage drilled at the end of year t,

and a, α, k_1 and k_2 were positive constants.

In Equation 4.5.6, $a < 1$, since $y = q - r$, $r \geq o$, and $q < w$. However, in Equation 4.5.7, $\alpha < 1$, because $y \leq R$ but no limitation on the magnitude of f has been imposed.

In order to estimate the relative predictive value of q, w, and f in determining y, the following multivariate regression model was proposed:

$$\ell n\, y = \beta_0 + \beta_1\, \ell n\, q + \beta_2\, \ell n\, w + \beta_3\, \ell n\, f \tag{4.5.8}$$

where β_i = the unknown population regression coefficients.

The equation is linear in the transformed variates, and thus the coefficients $\hat{\beta}_i$ can be estimated by means of linear regression.

Another way of expressing Equation 4.5.8 is

$$y = k\, q^c w^a f^d \tag{4.5.9}$$

with $a = \beta_2$, $c = \beta_1$, $d = \beta_3$, and $\ell n\, k = \beta_0$.

Similarly, if statistics on the cumulative number of wells drilled is used instead of cumulative well footage, then we could put

$$\ell n\, y = \beta_0 + \beta_1\, \ell n\, q + \beta_2\, \ell n\, w + \beta_3\, \ell n\, \Psi \tag{4.5.10}$$

where $\Psi(t)$ = the (all time) cumulative number of wells drilled up to and including the year t.

$= \Psi$, for convenience.

Also, let $\ell n\, y = \beta_0 + \beta_1\, \ell n\, q + \beta_2\, \ell n\, \Psi$, (4.5.11)

if data on oil in place were not available, or were not used.

This latter model would be of use if data were available on the cumulative number of wells drilled, but not on the cumulative well footage. Equations 4.5.8 and 4.5.10 provide a means for comparing the relative merits of q, w, f (and Ψ) to predict cumulative production.

If the hypothesis is made that the amount of oil in place discovered per footage of wells drilled depends upon the amount of oil in place left to be discovered then

$$w(f) = W(1-e^{cf}), \qquad (4.5.12)$$

provided $w(f) = o$, when $f = o$, as in equations 4.2.6 and 4.3.9 and where

$f(t)$ = cumulative footage of wells drilled at the end of year t,

$w(f)$ = cumulative discoveries of oil in place for f cumulative footage of wells drilled, and

W = ultimate resource of oil in place.

As wells are drilled each year, new pools of oil are found and defined, leading to new reserve estimates, and finally resulting in production. The relationship between production and well footage drilled which was implied by a pair of equations described previously was

$$w = k\,y^{1/a}, \qquad (4.5.6)$$

and

$$w = W\,(1-e^{cf}) \qquad (4.5.12)$$

Elimination of w results in a generalized form of the exponential equation.

A simpler relationship can be hypothesized by assuming that cumulative discoveries of oil in place are reflected in cumulative production within a period of say three to seven years. This proposal would result in an equation of the form

$$y\,(t+\tau) = R\,(1 - e^{cf(t)}) \qquad (4.5.13)$$

where τ = a time lag of say five years. This non linear equation could not be simplified further, because well footage was not a linear function of time. If cumulative well footage statistics were not available and data on the cumulative number of wells drilled were available, then Equation 4.5.13 could be modified to read

$$y\,(t+5) = R\,(1 - e^{c\Psi(t)}) \qquad (4.5.14)$$

While a great number of other models could also have been proposed, it was decided to limit the models hypothesized to those given in this Chapter. The results obtained from these might provide the basis for more sophisticated models, but this step was beyond the scope of the present study.

CHAPTER V

MATHEMATICAL ANALYSIS OF THE MODELS

SECTION ONE - THE MODELS SELECTED

Mathematical models may be divided into three general classes:

1) purely deterministic,
2) static or deterministic with simple random components, and
3) stochastic.[1]

As outlined in the previous Chapter, the general hypothesis of this study is that a reasonably simple static model (or models) exist which can describe the time pattern of the utilization of oil in the continental United States (excluding Alaska).

By hypothesizing any particular model M, an assumption $H_0(M)$ is implicitly made that a time pattern exists. The alternative hypothesis $H_1(M)$ is that no such time pattern exists, and that any time dependent analysis with M is invalid.

One criterion selected for testing the relative merits of each model was the distribution of the residuals. It was assumed that if a model was adequately describing the time pattern, then the residuals should be a random sequence in time. This (and other tests of the models selected) is described in the next section.

For such an analysis of the time pattern of events to have any validity for long term projections, it was necessary to use measurements which:

1) encompassed a fairly long span in time, and
2) would be reasonably consistent estimators of the long term utilization of oil, as t became large.

Some variables which were believed to satisfy these two criteria were as follows:

$\Delta y(t)$ = annual production, t=1859 to 1968.

$y(t)$ = cumulative production, t = 1859 to 1968.

$r(t)$ = proved reserves, t = 1920 to 1968.

$q(t)$ = cumulative discoveries, t = 1860 to 1968.

$v(t)$ = original oil in place by year of discovery, $t = 1920$ to 1968.

$w(t)$ = cumulative discoveries of oil in place, $t =$ 1919 to 1968.

$\Delta\Psi(t)$ = total wells completed per year, $t = 1859$ to 1968.

$\Psi(t)$ = cumulative number of wells drilled from 1859 to the year t, inclusive.

$\Delta f(t)$ = well footage drilled per year, $t = 1934$ to 1968.

$f(t)$ = estimated cumulative well footage drilled up to and including the year t, $t = 1934$ to 1968,
= f, for convenience.

$D(t)$ = domestic demand for all refined products in the United States (including Alaska), 1919 to 1968.

$G(t)$ = domestic demand for all gasoline in the United States (including Alaska), 1919 to 1968.

$H(t)$ = crude oil production in the Western Hemisphere,

$N(t)$ = crude oil production in North America,

The data for the above variables were restricted to the United States (and excluding Alaska whenever possible) and including any data on offshore activities in drilling, discoveries and production. The data base for the above variables, along with references to the sources of information, has been summarized in the Appendices. Data for annual production were also included for North America, South America, and World, as well as for various regions in North America (Canada, Alberta, Alaska, Mexico, and Cuba).

In order to use $f(t)$ in an equation, an estimate of cumulative footage drilled up to 1933 had to be made. There had been 862,665 wells drilled during the period 1859 to 1933. Assuming the average footage for these wells to be 2500 feet per well, cumulative footage of wells drilled up to and including 1933 would therefore be roughly 2500 x 862,665 feet. This is merely a guess based on the fact that the average depth of wells during 1927 - 1933 was 2947 feet, while the average footage for 1934 - 1968 was 3813 feet per well.[2,3]

Annual figures were available for only three variables (production, proved discoveries, and wells drilled) from 1859 onward, while observations for most of the other variables were available from 1920 onward. For this reason, some models were tested using data from a much shorter time period. The models selected in this study are summarized in Table 5.1.1 .

If a model was expressed as a function of time only, then the residuals could be calculated for $t=1,2,3,\ldots.T$, where T was the number of years for which annual data were available. If the same function was expressed in incremental form, time could be removed from the expression because Δt was always equal to one year. In this case it was assumed that annual data on the rate of change of a variable could be used to approximate the rate of change as expressed by the differentiated equation; that is, we equate

$$\Delta y = \dot{y}, \tag{5.1.1}$$

and

$$\Delta q = \dot{q}, \tag{5.1.2}$$

for these models.

This meant that slightly different hypotheses were being made for

1) the time dependent function, and
2) the same function in incremental form.

The parameters determined by the two approaches might therefore be different because of the approximation used above. It was assumed that if the data covered a large enough span of years, ($T > 40$, say) then the differences resulting from this effect should be small, and that the results from the two approaches could be compared.

If there were gross differences in the parameters as obtained from a time dependent function, and from the same function in incremental form, then some doubt could be cast on the adequacy of these functions to describe the data. The differential equation is a direct consequence of the original function; the proposal of the time dependent function as being descriptive of the nature of the time path is nearly equilavent to the proposal of the rate equation (using annual data) to describe the time path.

TABLE 5.1.1

THE MODELS SELECTED TO DESCRIBE THE LONG TERM UTILIZATION OF OIL

Model	Equation	Assumed Relationship	Variables	Parameters
Exponential Growth:				
M1	4.2.3	$\ln y = \ln k + ct$	$y; t$	$\ln k, c$
M2	4.3.1	$y(t) = a\,y(t-1) + b$	$y(t)$	a, b
Second Order Difference Equation:				
M3	4.3.11	$y(t) = c - ay(t-1) - by(t-2)$	$y(t)$	a, b, c
Polynomial in Time:				
M4	4.3.24	$y(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4$	$y; t$	β_i
Multivariate Models:				
M5	4.4.1	$\Delta y(t) = \beta_0 + \beta_1 D(t) + \beta_2 G(t) + \beta_3 H(t) + \beta_4 N(t) + \beta_5 r(t)$	$\Delta y; D, G, H, N, r$ *	β_i
M6	4.4.2	$\Delta y(t) = \beta_0 + \beta_1 [D(t) - \Delta y(t)] + \beta_2 [H(t) - \Delta y(t)] + \beta_3 [N(t) - \Delta y(t)] + \beta_4 r(t)$	$\Delta y; D, H, N, r$ *	β_i
M7	4.5.8	$\ln y = \beta_0 + \beta_1 \ln q + \beta_2 \ln w + \beta_3 \ln f$	$y; q, w, f$ *	β_i
M8	4.5.11	$\ln y = \beta_0 + \beta_1 \ln q + \beta_2 \ln \Psi$	$y; q, \Psi$	β_i
The Gompertz Curve:				
M9	4.2.7	$y = Rc^{d^t}$	$y; t$	R, c, d **
M10	4.2.10 5.1.1	$\Delta y = -ay \ln(y/R)$	$\Delta y; y$	a, R **
M11	4.2.17 4.2.18	$\ln[y(t+1)] = (1-d)(\ln R) + d \ln[y(t)]$	$y(t)$	$(1-d)(\ln R), d$

Model	Equation	Assumed Relationship	Variables	Parameters
The Logistic Curve:				
M 12	4.2.19***	$y = R/(1 + Ab^{t})$	$y;t$	R, A, b**
M 13	4.2.22*** 5.1.1	$\Delta y = cy(R - y)$	$\Delta y; y$	c, R**
M 14	4.2.24***	$1/y(t+1) = b/y(t) + (1-b)/R$	$y(t)$	$b, (1-b)/R$
The Generalized Logistic Curve:				
M 15	4.2.32	$y = R/(1 + Ae^{-at})^{\theta}$	$y;t$	R, A, a, θ**
M 16	4.2.33 4.5.1	$\Delta y = ky[1 - (y/R)^{\beta}]$	$\Delta y; y$	k, R, β**
The Beta Function:				
M 17	4.2.34	$\Delta y = ky^{m}(R-y)^{n}$	$\Delta y; y$	k, m, n, R**
M 18	4.2.36 5.1.1	$\Delta y = ky^{m}(R-y)^{1-m}$	$\Delta y; y$	k, m, R**
Miscellaneous Relationships:				
M 19	4.5.1***	$y = p.r./m - b/m$	$y; q, w$*	m, b; $R < (1-b)/m$
M 20	4.5.3	$\ln y = c \ln q + \ln b$	$y;q$	$c, \ln b$; $\ln b = (1-c) \ln R$
M 21	4.5.4	$y = bq^{c}$	$y;q$	b, c;** $\ln b = (1-c) \ln R$
M 22	4.5.13	$y(t) = R(1 - e^{c\Psi})$	$y;\Psi$	R, c**

* Data not available for the period 1859 to 1968.

** A computer algorithm to minimize the sum of squares of the residuals would be necessary to calculate the parameters; the parameters in all the other equations could be estimated by means of linear regression.

*** Production rather than discoveries have been used as the dependent variable in the expression to allow comparisons between models; q (or Δ q) could be exchanged for y(or Δy) in most models if warranted.

Similarly, a function could be hypothesized in difference equation form, resulting in slightly different estimates of the parameters than would be obtained if the function were expressed in differential equation form. Again, if a long enough span of time T is used, then this kind of error should be small.

In some models, transformations of variates have been made in order to reduce the equation to linear form. These transformations included the use of logarithms and reciprocals of the dependent variable. Data used in such models were transformed prior to carrying out the regression between variates. In this case, the relative magnitude of the residuals was affected by the transformation. The residuals being minimized by least squares were in a non-linear form.

While study of the analytical portion of the model was simplified by the transformation, the analysis of the stochastic portion of the model became more difficult under the assumptions that have been made in the previous Chapter regarding the nature of the residuals. This might present a problem where there was a wide range of values for the variates over the period being studied, such as from nearly zero to some large number. In any case, if a transformation led to a much different result than obtained when the data were not transformed, then one of the approaches should be rejected. The selection of the most appropriate model would be dependent upon the tests made on all models, including the time pattern of the residuals.

Each statistical analysis was thus to be built on a model proposed to link observable reality with the basic mechanism generating the observations. The model should be a parsimonious description of nature; its functional form should be simple, and the number of parameters and components should be a minimum. The model should be structured in such a way that each parameter can be interpreted easily and identified with some aspect of reality. The functional form should be sufficiently tractable to permit the sort of mathematical manipulations required for the estimation of its parameters and other inferences about its nature.

REFERENCES:

1. Donald F. Morrison, Multivariate Statistical Methods, McGraw-Hill Book Company, New York, 1967, p. 1.

2. H. J. Struth, Editor, The Petroleum Data Book, The Petroleum Engineer Publishing Co., Dallas, Texas, 1947, p. D-40.

3. See Appendix C and Appendix D.

SECTION TWO - TESTS OF THE MODELS

With a knowledge of the general characteristics of various models, and of the results that had been obtained using some of the models, tests which would assess the relative merits of each model had to be proposed. In proposing criteria, or tests, of the models, a distinction was made between what this writer believed were:

1. the necessary properties of the model, and of its parameters, and
2. the desirable properties of the model, and of its parameters.

In other words, more weight, or credence, was assigned to some of the tests, or criteria.

Let $z(t) = z =$ the actual (untransformed) observations of the variate z, for which a time pattern was assumed to exist $t = 1,2,3,\ldots.$

$z'(t) = z' =$ the values calculated for the variate z using a particular model, $t = 1,2,3,\ldots.$

Also, put $$e(t) = z'(t) - z(t) \tag{5.2.1}$$

= the residuals, or difference between the calculated values and the observed values, $t = 1,2,3,\ldots N$,

where N = the number of years for which data was available.

Then some of the necessary properties of the models were assumed to be:

A. The expression for $z = y$ (or q) would be a monotonically increasing function of time.

B. The functional relationship of the model for $z = y$ (or q, or Δy, or Δq) could be expressed in a form suitable for reasonably simple calculations of the time pattern.

C. The variance of the residuals would be minimized.

D. The model should be unbiased. (ie. - The residuals $e(t)$ should have an expectation equal to zero.)

E. A single maximum exists in those models with $z = \Delta y$ (or Δq).

F. $o \leq y \leq q \leq R$.

G. If $y = R$, then $q = R$, and $\dot{y} = \dot{q} = o$.

H. A model using $z = q$ should result in essentially the same long term estimate for R as this model would give using $z = y$, because y lags q by about 10 to 15 years.

I. The estimate of the parameter R should probably be greater than 193 billion barrels for the United States; this figure was 50% of the original oil in place discovered up to 1968.[1]

J. The estimate of the parameter R should be less than 2319 billion barrels for the United States, which is six times the amount of oil in place found up to 1968. (This figure is greater than any estimate of the ultimate resource of oil in place, and actual cumulative production would be less.)

K. If a model for y(or q) could also be expressed in differential equation form, then this latter model with $\Delta y = \dot{y}$ (or $\Delta q = \dot{q}$) as in Equation 5.1.1, should result in an estimate for R consistent with the estimate for R obtained using y (or q) in the integrated form.

Some of the features believed to be desirable for each model were:

L. The residuals e(t) (using untransformed variates) should be normally distributed.

M. The model should exhibit the feature of additivity if used for several regions, and then for the total of the regions.

N. The parameters should have some meaningful interpretation with some aspect of reality.

O. The model should be invariant with change in scale of the variables. For example, the estimate of the parameter R should be the same whether the variables are measured in thousands of barrels, or in millions of barrels.

P. The model should be in dimensionally balanced form. (This implies that homogeneous functions were preferable to non-homogeneous functions, if each variable in the function were measured in the same units, such as barrels.)[2]

Q. Annual data for the variables should be available for a period of at least 40 years, because less than one half of the resource has been produced up to 1968.[3]

The analytic features of each model were fully discussed in the preceding Chapter. The test statistics for the stochastic portion of the model are given in Equations 5.2.2 to 5.2.16.

One obvious measure of the adequacy of each model is given by the coefficient of correlation,

$$\rho = \mathrm{cov}(z',z)\left(\mathrm{var}(z')\,\mathrm{var}(z)\right)^{-\frac{1}{2}} \qquad (5.2.2)$$

= the correlation coefficient of the observed values (z) with the calculated values (z'),

where var (z') = the variance of z',

var (z) = the variance of z, and

cov (z', z) = the covariance of z' and z.

The relative significance of ρ can then be assessed by the formula

$$\mathrm{Prob}\ (\rho_l \leq \rho \leq \rho_u) = (1-\alpha) \qquad (5.2.3)$$

where ρ_l = the lower limit of ρ,

ρ_u = the upper limit of ρ, and

α = the level of significance expressed in decimal form. (If each variable increased with time, the fact that $\rho > 0$ may not be significant.)

In multivariate linear models, the relative significance of the multiple correlation coefficient could be assessed. Here, one degree of freedom would be lost for each independent variable added. The relative significance of each independent variable in prediction could be measured by the estimated value and standard error of the regression coefficients.[4]

The magnitude of the deviations is given by:

$$S = \sum_{t=1}^{N} e(t) \,, \tag{5.2.4}$$

and
$$SA = \sum_{t=1}^{N} | e(t) | \tag{5.2.5}$$

Where $e(t) = z'(t) - z(t)$, or $e = z' - z$, for convenience.

= the residuals, or the differences between the calculated values and the observed values.

A more appropriate measure of the residuals is given by the sum of squares of the residuals, SSE,

where
$$SSE = \sum_{t=1}^{N} [e(t)]^2 \tag{5.2.6}$$

In equation 5.2.4 the expected value of S is zero if SSE is minimized. However, there is no guarantee that S=0 if ordinary least squares is carried out using transformed variates.

An unbiased estimator of the variance of the residuals, var(e), is s,

with
$$s^2 = SSE/(n-p), \tag{5.2.7}$$

where s = the standard error of estimate,

$n = N-1$, and p = the number of independent variables.

The serial correlation of the population of residuals is given by:

$$\rho_s = \text{cov}\,[\,e(t+1),\ e(t)\,]\,[\text{var}\ e(t+1)\ \text{var}\ e(t)\,]^{-\frac{1}{2}} \tag{5.2.8}$$

The sample statistic r_s (which is an estimate for ρ_s) can be calculated from the correlation between successive residual terms. If the residuals are divided into two series

$$S_1 = \{ e(1),\ e(2),\ e(3),\ \ldots\ e(N-1) \} \tag{5.2.9}$$

and
$$S_2 = \{ e(2),\ e(3),\ e(4),\ \ldots e(N) \},$$

then r_s = the simple correlation coefficient between S_1 and S_2.

A value of r_s significantly different than zero would indicate that a time pattern existed for the residuals, and that the model was not adequately describing the time pattern of the rate of utilization. Unfortunately, the test for significance could not be used in the same way as for the usual correlation coefficient since it is necessay to make allowances for the regression relationship from which the residuals were derived, as in the case of lagged equations or difference equations.

In a multivariate linear model, if

$$z_i = \hat{\beta}_0 + \hat{\beta}_1 X i_1 + \hat{\beta}_2 X i_2 + \ldots + \hat{\beta}_p Xip + e_i$$
$$= z'_i + e_i \tag{5.2.10}$$

where $\hat{\beta}_j$ = the estimates of the population regression coefficients, $j = 1, 2, 3, \ldots p$

p = the number of independent variables,

N = the number of observations, and $i = 1, 2, 3, \ldots N$,

then the estimate of multiple correlation R is given by

$$R^2 = SSR/SST \tag{5.2.11}$$

where SSR = the sum of products in regression

$$= \sum_{j=1}^{p} \hat{\beta}_j \left[\sum_{i=1}^{N} X ij\, z_i - \frac{1}{N} \sum_{i=1}^{N} X ij \sum_{i=1}^{N} z_i \right], \tag{5.2.12}$$

and SST = the total sum of squares for the dependent variable

$$= \sum z_i^2 - \frac{1}{N} (\sum z_i)^2 \tag{5.2.13}$$

In this case we write

$\sum e_i^2$ = the sum of squares of the residuals

$$= \sum (z_i - z'_i)^2$$
$$= SST - SSR = SSE \tag{5.2.14}$$

Also, the quantity $F = MSR/MSE$ (5.2.15)

has an F distribution with p and n-p degrees of freedom

where MSR = the mean square for regression

$$= SSR/p, \tag{5.2.16}$$

and MSE = the mean square for error

$$= SSE/(n-p) \tag{5.2.17}$$

The results of the above computations are usually displayed in an analysis of variance table, such as below:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square	F
Regression	p	SSR	MSR	MSR/MSE
Error	n-p	SSE	MSE	
Total	n	SST		

If the residuals are independent, and normally distributed, it may be shown that β_i is normally distributed with mean β_i and an unbiased estimate of variance

$$s_i^2 = s^2\, C_{ii} \qquad (5.2.18)$$

where Cii is the diagonal element of the variance covariance matrix C for the model in Equation 5.2.10.

Also $$t_i = (\hat{\beta}_i - \beta_i)/s_i \qquad (5.2.19)$$

has a Student's t-distribution with n-p degrees of freedom.

Equations 5.2.18 and 5.2.19 can give an indication of whether the regression coefficients are significant under the hypothesis that

$$\beta_i = 0, \quad i = 1, 2, 3, \ldots p$$

A table such as the following would indicate the information required for regression coefficients in Equation 5.2.10

Type of Estimate	$i = 0$	$i = 1$	$i = 2$ . . . .
β_1	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
s_i		s_1	s_2
t_i		t_1	t_2

It is also possible to augment the analysis of variance table in such a way as to sort out only those variables Xip accounting for a significant level of the variation in the dependent variable. However, a stepwise regression algorithm provided substantially the required results for the purpose of this study.

We let Xa = the independent variable which accounts for the largest proportion of variation,

Xb = the variable which accounts for the largest proportion of the remaining variation,

and so on for the remaining $p-2$ independent variables,

also, let r_a = the partial correlation of z and Xa

$r_{b.a}$ = the partial correlation of z and Xb with the effect of Xa removed

$r_{c.cb}$ = the partial correlation of z and Xc with the effect of Xa and Xb removed.

Then $V_a = r_a^2$ = the proportion of variation accounted for by Xa

$V_b = (1-r_a^2)\ r_{b.a}^2$

= the proportion of variation accounted for by X_b

$V_c = (1-r_a^2)(1-r_{b.a}^2)\ r_{c.ab}^2$

= the proportion of variation accounted for by X_c

and V_i = the proportion of variation in the dependent variable accounted for by the independent variable X_i, when the X_i's are entered in the stepwise order. V_i would ordinarily be expressed in percent, as indicated in the following layout.

PROPORTION OF VARIATION IN THE DEPENDENT VARIABLE ACCOUNTED FOR BY EACH INDEPENDENT VARIABLE:

Independent Variable		Estimate of the Proportion Accounted for (V_j)
X_a	$i=a$	V_a (in percent)
X_b	$i=b$	V_b (in percent)
X_c	$i=c$	V_c (in percent)
..... and so on for the remaining p-3 variables.		

The above calculations could be conveniently handled by means of International Business Machines A Programming Language (APL/360).[5,6,7]

If a difference equation is not used, the Durbin - Watson statistic provides a convenient measure of the serial correlation of the residuals, with:

$$dw = \sum_{t=1}^{N} [e(t) - e(t-1)]^2 / SSE, \qquad (5.2.21)$$

= the Durbin - Watson statistic.

Tables of the upper and lower significance points for dw are given in the 1951 article by Durbin and Watson. A low value for dw would indicate positive serial correlation. The test does not apply for lagged equations, or for the case of regression through the origin.[8,9,10]

A nonparametric measure of the nature of the residuals is given by:

$$P(m \leq \mu) = \alpha \qquad (5.2.22)$$

= the probability that the observed number of runs is less than μ,

where m = the observed number of runs,

p = the observed number of positive signs,

n = the observed number negative signs,

μ = the mean number of runs,

$= 1 + 2\,pn / (p+n)$

σ = the standard deviation of the number of runs,

and $\sigma^2 = 2\,pn\,(2pn - p - n) / (p+n)^2\,(p+n-1)$

Usually $\alpha = 0.025$ for a 95% confidence level.[11,12]

Besides the above tests, it was of interest to determine the year of the maximum residual size. If the maximum deviation occurred at the very beginning (or end) of the time series, one could suspect (but not prove) that the model tended to deviate from the true time pattern of the observation.

A much more powerful test of such a tendency in the model is given in Chapter IV by:

$$e(t) = \beta_0 + \beta_1 t + \beta_2 t^2 \qquad (4.3.25)$$

If $\beta_2 > 0$, then the residuals would be concave upward. The reduction of the mean square for error by introducing the term $\beta_2 t^2$ would also indicate the significance of a simple (parabolic) time trend.

An indication of the normality of the residuals is given by the coefficients of skewness and kurtosis of the residuals, which are defined by:

$$Sk = k_3 / k_2^{3/2} , \qquad (5.2.23)$$

and $$Ku = k_4 / k_2^2 \qquad (5.2.24)$$

where $k_2 = Nm_2 \div (N-1)$

$$k_3 = N^2 m_3 \div (N-1)(N-2)$$

$$k_4 = \frac{N^2 ((N+1)m_4 - 3(N-1)m_2^2)}{(N-1)(N-2)(N-3)}$$

$$m_2 = S_2 - S_1^2$$

$$m_3 = S_3 - 3S_1 S_2 + 2S_1^3$$

$$m_4 = S_4 - 4S_1 S_3 + 6S_2 S_1^2 - 3S_1^4$$

and $$S_i = \frac{1}{N} \sum_{t=1}^{N} (e(t))^i , \; i = 1,2,3,4.$$

By expanding, we arrive at

$$Sk = \frac{N^2 (S_3 - 3S_1 S_2 + 2S_1^3)}{(N-1)(N-2)} \left\{ \frac{N-1}{N(S_2 - S_1^2)} \right\}^{3/2} \qquad (5.2.25)$$

and $$Ku = (N-1) ((N+1) (S_4 - 4S_1 S_3 + 6S_2 S_1^2 - 3S_1^4)$$

$$- 3(N-1) (S_2 - S_1^2)^2) \div (N-2)(N-3) (S_2 - S_1^2)^2 \qquad (5.2.26)$$

If the residuals are normally distributed, the coefficients of skewness and kurtosis have standard deviations of

$$\sigma_1 = \left\{ \frac{6N(N-1)}{(N-2)(N+1)(N+3)} \right\}^{\frac{1}{2}} \qquad (5.2.27)$$

$$\doteq (6/n)^{\frac{1}{2}} \quad \text{for skewness}$$

and $$\sigma_2 = \left\{ \frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)} \right\}^{\frac{1}{2}} \qquad (5.2.28)$$

$$\doteq (24/n)^{\frac{1}{2}} \quad \text{for kurtosis}$$

This is essentially the form in which calculations were carried out, using the APL/360 system.[13]

REFERENCES:

1. See Appendix A and Appendix B.

2. G. James and R. C. James, Mathematics Dictionary, Multilingual Edition, D. Van Nostrand Company, Inc., Princeton, New Jersey, 1959, p. 188.

3. See Appendix A.

4. E. S. Pearson and H. O. Hartley, Editors, Biometrika Tables for Statisticians, Volume I, Cambridge University Press, 1962, pp. 30 - 32, p. 141.

5. K. W. Smillie, An Introduction to Regression and Correlation, The Ryerson Press, Toronto, 1966, pp. 42,47,48,58,62,65.

6. K. W. Smillie, STATPACK2: An APL Statistical Package, Second Edition, Publication No. 17, Department of Computing Science, University of Alberta, Edmonton, Alberta, February, 1969, p. 33.

7. E. S. Keeping, Introduction to Statistical Inference, D. Van Nostrand Company, Inc., Princeton, New Jersey, 1962, p. 292.

8. Smillie, An Introduction....p. 93.

9. J. Durbin and G. S. Watson, Testing for Serial Correlation in Least Squares Regression.I, Biometrika, 37, 1950, pp. 409 - 428.

10. J. Durbin and G. S. Watson, Testing for Serial Correlation in Least Squares Regression.II, Biometrika, 38, 1951, pp. 159 - 178.

11. Smillie, An Introduction....pp. 93 - 96.

12. Pearson and Hartley, pp. 104 - 110.

13. See Appendix: "Computer Programs Used".

SECTION THREE - ANALYSIS OF RESULTS

The models M1 to M22 shown in Table 5.1.1 all have production (or some transformation of production) as the dependent variable. This allowed comparison of the results obtained from the various models.

Data for three of the models (M5, M6, and M19) were available only for 1920 to 1968. In one model (M7) data were available for only 1934 to 1968. In all other models, data on production and on wells drilled were available for the years 1859 through to 1968; in three models (M8, M20 and M21), estimates of cumulative discoveries were available for the years 1860 to 1968. The series of estimates for cumulative discoveries for the year ending 1969 were not available in time to be incorporated in this study.

It was decided to study the test results from the multivariate models first, so as to get some idea of the relative contribution of various variables as estimators of production. These statistics have been summarized in Tables 5.3.1, 5.3.2, 5.3.5, 5.3.4, 5.3.5, and 5.3.6. All figures in these (and subsequent) tables have been rounded to four figure accuracy. Note that a number followed by E-k indicates that the number is to be multiplied by 10^{-k}.

A variety of statistics were calculated concerning the residuals in each model. The residuals in each of the multivariate models were not random, so far as the expected number of runs were concerned. This may have been caused by periodic highs or lows in production, which could have some basis apart from any long term trend. However, the residuals in models M3 and M8 were significantly skewed and peaked, as measured by the tests of skewness and kurtosis, and so were not normal in distribution.

The fact that residuals in all the twenty-two models did not have a random number of runs tended to confirm the idea that some (independent) cycle of highs and lows for production indeed existed. The t-values for this test are shown in Tables 5.3.8, 5.3.20, 5.3.23 and 5.3.26. In model M3, the level of significance was only valid at the 2% level. However, when model M3 was expressed with y as

continued on p.128

TABLE 5.3.1

RESULTS FROM MULTIVARIATE MODEL M7 FOR CUMULATIVE PRODUCTION USING DATA FROM 1934 TO 1968

I FUNCTIONAL FORM OF M7:

$$\ln y = \beta_0 + \beta_1 \ln q + \beta_2 \ln w + \beta_3 \ln f$$

II ANALYSIS OF VARIANCE:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square	F
Regression	3	8.586	2.862	5479
Error	31	1.619E-2	5.223 E-4	
Total	34	8.602		

III MEASURES OF THE SAMPLE REGRESSION COEFFICIENTS

Type of Estimate	$i=0$	$i=1$	$i=2$	$i=3$
$\hat{\beta}$	-3.038	0.9564	-0.2035	0.3702
s_i		0.2301	0.2629	0.1245
t_i		4.156	-0.7743	2.974
α_i		$<0.05\%$	$>5\%$	$<0.5\%$

IV PROPORTION OF VARIATION IN THE DEPENDENT VARIABLE ACCOUNTED FOR BY EACH INDEPENDENT VARIABLE:

Independendent Variable	Estimate of the Proportion Accounted for (Vj)
$\ln q$	99.4924%
$\ln w$	0.0036%
$\ln f$	0.3157%

V MULTIPLE CORRELATION COEFFICIENT:

$R^2 = 0.9981$

VI STANDARD ERROR OF THE RESIDUALS:

$s = 2.285 \times 10^{-2}$

TABLE 5.3.2

RESULTS FROM MULTIVARIATE MODEL M5 FOR ANNUAL PRODUCTION USING DATA FROM 1920 TO 1968

I FUNCTIONAL FORM OF M5:

$$\Delta y(t) = \beta_0 + \beta_1 D(t) + \beta_2 G(t) + \beta_3 H(t) + \beta_4 N(t) + \beta_5 r(t)$$

II ANALYSIS OF VARIANCE:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square	F
Regression	5	2.919E7	5.822E6	3115
Error	43	8.035E4	1.869E3	
Total	48	2.919E7		

III MEASURES OF THE SAMPLE REGRESSION COEFFICIENTS:

Type of Estimate	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$\hat{\beta}_i$	-1.544E2	-3.419E-1	3.832E-1	-2.920E-1	1.318	2.179E-2
s_i		1.262E-1	2.359E-1	1.220E-1	1.282E-1	4.205E-3
t_i		-2.708	1.624	-2.393	1.028E1	5.182
α_i		$<0.5\%$	$>5\%$	$<2.5\%$	$<0.05\%$	$<0.05\%$

IV PROPORTION OF VARIATION IN THE DEPENDENT VARIABLE ACCOUNTED FOR BY EACH INDEPENDENT VARIABLE:

Independent Variable	Estimate of the Proportion Accounted for (V_j)
N(t)	98.2256 %
r(t)	1.1518 %
D(t)	0.3077 %
H(t)	0.0227 %
G(t)	0.0169 %

V MULTIPLE CORRELATION COEFFICIENT:

$R^2 = 0.9972$

VI STANDARD ERROR OF THE RESIDUALS:

$s = 43.23$

TABLE 5.3.3

RESULTS FROM MULTIVARIATE MODEL M6 FOR ANNUAL PRODUCTION USING DATA FROM 1920 TO 1968

I FUNCTIONAL FORM OF M6

$$\Delta y(t) = \beta_0 + \beta_1 [D(t) - \Delta y(t)] + \beta_2 [H(t) - \Delta y(t)] + \beta_3 [N(t) - \Delta y(t)] + \beta_4 r(t)$$

II ANALYSIS OF VARIANCE:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square	F
Regression	4	2.868E7	7.171E6	625.6
Error	44	5.044E5	1.146E4	
Total	48	2.919E7		

III MEASURES OF THE SAMPLE REGRESSION COEFFICIENTS:

Type of Estimate	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$
β_i	-5.801E1	-8.406E-1	8.673E-1	3.432E-1	6.185E-2
s_i		2.300E-1	2.525E-1	5.050E-1	7.262E-3
t_i		-3.655	3.435	6.795E-1	8.516
α_i		<0.05%	<0.05%	>5%	<0.05%

IV PROPORTION OF VARIATION IN THE DEPENDENT VARIABLE ACCOUNTED FOR BY EACH INDEPENDENT VARIABLE:

Independent Variable	Estimate of the Proportion Accounted (V_j)
$r(t)$	95.6777 %
$H(t) - \Delta y(t)$	2.0698 %
$D(t) - \Delta y(t)$	0.5164 %
$N(t) - \Delta y(t)$	0.0018 %

V MULTIPLE CORRELATION COEFFICIENT:

$R^2 = 0.9827$

VI STANDARD ERROR OF THE RESIDUALS:

$s = 107.1$

TABLE 5.3.4

RESULTS FROM MULTIVARIATE MODEL M4 FOR CUMULATIVE PRODUCTION

I FUNCTIONAL FORM OF M4:

$$y(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4$$

II ANALYSIS OF VARIANCE:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square	F
Regression	4	6.183E10	1.546E10	72660
Error	105	2.234E7	2.127E5	
Total	109	6.186E10		

III MEASURES OF THE SAMPLE REGRESSION COEFFICIENTS:

Type of Estimate	i = 0	i = 1	i = 2	i = 3	i = 4
β_i	-956.9	198.3	-8.234	9.075E-2	3.135E-4
s_i		28.74	1.046	1.413E-2	6.316E-5
t_i		6.900	-7.871	6.422	4.965
α_i		< 0.05%	< 0.05%	< 0.05%	< 0.05%

IV PROPORTION OF VARIATION IN THE DEPENDENT VARIABLE ACCOUNTED BY EACH INDEPENDENT VARIABLE:

Independent Variable	Estimate of the Proportion Accounted for (Vi)

(An estimate of the proportion of variation accounted for by each variable could not readily be calculated on a 110 x 5 data matrix using standard computer algorithms on the APL/360 system. These estimates are therefore not given for model M4.)

V MULTIPLE CORRELATION COEFFICIENT:

$$R^2 = 0.9996$$

VI STANDARD ERROR OF THE RESIDUALS:

$$s = 461.2$$

TABLE 5.3.5

RESULTS FROM MULTIVARIATE MODEL M8 FOR CUMULATIVE PRODUCTION

I FUNCTIONAL FORM OF M8:

$$\ln y(t) = \beta_0 + \beta_1 \ln q(t) + \beta_2 \ln \Psi(t)$$

II ANALYSIS OF VARIANCE:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square	F
Regression	2	849.7	424.8	13470
Error	106	3.342	3.153E-2	
Total	108	853.0		

III MEASURES OF THE SAMPLE REGRESSION COEFFICIENTS:

Type of Estimate	$i = 0$	$i = 1$	$i = 2$
β_i	-7.011	0.7716	0.6378
s_i		3.318E-2	3.113E-2
t_i		23.26	20.49
α_i		$<0.05\%$	$<0.05\%$

IV PROPORTION OF VARIATION IN THE DEPENDENT VARIABLE ACCOUNTED FOR BY EACH INDEPENDENT VARIABLE:

Independent Variable	Estimate of the Proportion Accounted for (Vi)
$\ln q(t)$	98.0562%
$\ln \Psi(t)$	1.5520%

V MULTIPLE CORRELATION COEFFICIENT:

$$R^2 = 0.9961$$

VI STANDARD ERROR OF THE RESIDUALS:

$$s = 0.1776$$

TABLE 5.3.6

RESULTS FROM MODEL M3 FOR CUMULATIVE PRODUCTION

I FUNCTIONAL FORM OF M3:

$$y(t) = c - a\,y(t-1) - b\,y(t-2)$$

II ANALYSIS OF VARIANCE:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square	F
Regression	2	6.130E10	3.065E10	7.899E6
Error	105	4.074E5	3.880E3	
Total	107	6.130E10		

III MEASURES OF THE SAMPLE REGRESSION COEFFICIENTS:

Type of Estimate	$i=0$	$i=1$	$i=2$
$\hat{\beta}_i$	1.147E1	2.020	-1.019
s_i		2.803E-2	2.917E-2
t_i		7.206E1	-3.495E1
α_i		<0.05%	<0.05%

VI PROPORTION OF VARIATION IN THE DEPENDENT VARIABLE ACCOUNTED FOR BY EACH INDEPENDENT VARIABLE:

Independent Variable	Estimate of the Proportion Accounted for (V_i)
y(t-1)	99.9916 %
y(t-2)	0.0077 %

V MULTIPLE CORRELATION COEFFICIENT:

$$R^2 = 0.9999$$

VI STANDARD ERROR OF THE RESIDUALS:

$$s = 62.29$$

TABLE 5.3.7

STATISTICS OF THE RESIDUALS IN MODELS M3, M4, M5, M6, M7 AND M8

STATISTIC	M3	M4	M5	M6	M7	M8
Number of residuals	108	110	49	49	35	109
Maximum deviation	-243.7	-1341	-110.3	227.7	-0.04566	-1.055
Year of maximum	1958	1968	1921	1968	1939	1860
Range of deviations	450.7	2524	202.9	432.1	0.08662	1.650
Median residual	-9.539	-18.51	2.504	-0.2594	-0.003605	0.02643
Coefficient of skewness	-0.3760	0.1852	-0.2819	-0.1172	0.08668	-1.542
t-value for skewness	-1.617	0.8038	-0.8296	-0.3447	0.2180	-6.661**
Coefficient of kurtosis	4.488	0.2675	0.1211	-0.6598	-0.2772	13.02
t-value for kurtosis	9.733**	0.5854	0.1813	-0.9876	-0.3564	28.37**
Dependent variable	y	y	Δy	Δy	$\ln y$	$\ln y$

* Significant at the 1% level.

** Significant at the 0.1% level.

TABLE 5.3.8

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN MODELS M3, M4, M5, M6, M7 AND M8

TEST STATISTICS	M3	M4	M5	M6	M7	M8
Durbin-Watson statistic	n.a.	0.0620	0.4543	0.4984	0.3745	0.6847
Number of runs of positive and negative signs:						
Observed runs, m	37	10	9	10	5	8
Positive signs, p	36	52	26	24	17	63
Negative signs, n	72	58	23	25	18	46
Population mean, μ	49	55.84	25.41	25.49	18.49	54.17
Standard deviation, σ	4.592	5.204	3.450	3.462	2.912	5.068
t-value, $(\mu - m)/\sigma$	2.613	8.807**	4.756**	4.474**	4.631**	9.110**
Parabolic regression of the residuals with time:						
t-value of $\hat{\beta}_2$	-1.052	n.a.	-2.339	-0.6348	-0.05270	2.109
F for regression	0.9574	n.a.	2.867	0.3424	0.03585	2.403

* Significant at the 1% level.

** Significant at the 0.1% level.

n.a. Test not applicable.

TABLE 5.3.9

RESULTS FROM MODEL M3†

I FUNCTIONAL FORM OF M3†:

$$y = \beta_0 + \beta_1 m_1^{\,t} + \beta_2 m_2^{\,t}$$

($m_1 = 1.024$; $m_2 = 0.9953$, rounded to four figures.)

II ANALYSIS OF VARIANCE:

Source of Variation	Degress of Freedom	Sums of Squares	Mean Square	F
Regression	2	6.153E10	3.077E10	10220
Error	107	3.220E8	3.009E6	
Total	109	6.186E10		

III MEASURES OF THE SAMPLE REGRESSION COEFFICIENTS:

Type of Estimate	i = 0	i = 1	i = 2
$\hat{\beta}_i$	-1.386E5	1.040E4	1.327E5
s_i		123.5	3825
t_i		84.20	34.71
α_i		<0.1%	<0.1%

IV MULTIPLE CORRELATION COEFFICIENT:

$R^2 = 0.9948$

V STANDARD ERROR OF THE RESIDUALS:

$s = 1735$

TABLE 5.3.10

STATISTICS OF THE RESIDUALS IN MODEL M3†

STATISTIC	
Number of residuals	110
Maximum deviation	-4171
Year of maximum	1859
Range of deviations	6612
Median residual	135.8
Coefficient of skewness	-0.2315
t-value for skewness	-1.005
Coefficient of kurtosis	-1.094
t-value for kurtosis	-2.393
Dependent variable	y

TABLE 5.3.11

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN MODEL M3†

TEST STATISTICS	
Durbin-Watson statistic	n.a.
Number of runs of positive and negative signs:	
Observed runs, m	4
Positive signs, p	57
Negative signs, n	53
Population mean, μ	55.93
Standard deviation, σ	5.213
t-value, $(\mu - m)/\sigma$	9.961**
Parabolic regression of the residuals with time:	
t-value of $\hat{\beta}_2$	-1.649
F for regression	1.370

** Significant at the 0.1% level.

n.a. The test is not applicable.

TABLE 5.3.12

CORRELATION MATRIX FOR MODEL M3

Variable			
$y(t)$	1	0.9999	0.9998
$y(t-1)$		1	0.9999
$y(t-2)$			1

TABLE 5.3.13

CORRELATION MATRIX FOR MODEL M4

Variable					
y	1	0.9984	0.9842	0.9421	0.8413
t		1	0.9688	0.9175	0.8672
t^2			1	0.9861	0.9585
t^3				1	0.9921
t^4					1

TABLE 5.3.14

CORRELATION MATRIX FOR MODEL M5

Variable						
$\Delta y(t)$	1	0.9799	0.9825	0.9818	0.9911	0.9781
D (t)		1	0.9968	0.9988	0.9950	0.9450
G (t)			1	0.9958	0.9926	0.9591
H (t)				1	0.9972	0.9448
N (t)					1	0.9547
r (t)						1

TABLE 5.3.15

CORRELATION MATRIX FOR MODEL M6

Variable					
$\Delta y(t)$	1	0.8791	0.9220	0.7792	0.9781
$D(t)-\Delta y(t)$		1	0.9895	0.9620	0.8791
$H(t)-\Delta y(t)$			1	0.9500	0.8701
$N(t)-\Delta y(t)$				1	0.6930
$r(t)-\Delta y(t)$					1

TABLE 5.3.16

CORRELATION MATRIX FOR MODEL M7

Variable				
y(t)	1	0.9975	0.9800	0.9932
q(t)		1	0.9898	0.9865
w(t)			1	0.9563
f(t)				1

TABLE 5.3.17

CORRELATION MATRIX FOR MODEL M8

Variable			
y(t)	1	0.9902	0.9880
q(t)		1	0.9645
Ψ(t)			1

TABLE 5.3.18

SUMMARY OF RESULTS FROM MODELS M1, M2, M11, M14, M19 AND M20

STATISTIC	M1	M2	M11	M14	M19	M20
Number of observations	110	109	109	109	49	109
Degress of freedom for error	108	107	107	107	47	107
Error sum of squares	124.2	5.170E6	21.55	0.1993	2.275E8	16.58
Standard error of the residuals	1.072	219.8	0.4488	4.316E-2	2200	0.3936
Correlation between observed and calculated values, r^2	0.8811**	0.9999**	0.9747**	0.9513**	0.9921**	0.9806**
Regression coefficients: $\hat{\beta}_0$	2.512	155.3	0.9268	1.168E-2	-55570	-5.034
$\hat{\beta}_1$	9.111E-2	1.040	0.8984	3.964E-3	466900	1.427
Standard error, $\hat{\sigma}$	3.220E-3	9.215E-4	1.398E-2	8.672E-5	6080	1.943
t-value, $\hat{\beta}_1 / \hat{\sigma}$	28.30**	1128.83**	64.24**	45.71**	76.79**	73.47**
Dependent variable	$\ln y$	y	$\ln y$	1/y	y	$\ln y$

** Significant at the 0.1% level.

TABLE 5.3.19

STATISTICS OF THE RESIDUALS IN MODELS M1, M2, M11, M14, M19 AND M20

STATISTIC	M1	M2	M11	M14	M19	M20
Number of residuals	110	109	109	109	49	109
Maximum deviation	-8.818	-483.6	3.967	0.3647	-7950	-2.826
Year of maximum	1859	1965	1860	1861	1929	1860
Range of deviations	9.528	925.4	4.453	0.3764	14450	3.337
Median residual	0.3367	-88.95	-0.0144	-0.01138	389.7	0.0703
Coefficient of skewness	-5.631	0.2791	6.522	6.548	-0.8020	-3.772
t-value for skewness	-24.44**	1.206	28.17**	28.29**	-2.360	-16.30**
Coefficient of kurtosis	43.27	-0.6837	57.97	50.40	4.163	24.70
t-value for kurtosis	94.68**	-1.490	126.3**	109.80**	6.231**	53.81**
Dependent variable	$\ell n\, y$	y	$\ell n\, y$	$1/y$	y	$\ell n\, y$

** Significant at the 0.1% level.

TABLE 5.3.20

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN MODELS M1, M2, M11, M14, M19 AND M20

TEST STATISTICS	M1	M2	M11	M14	M19	M20
Durbin-Watson statistic	0.2636	0.0848	0.5400	0.9019	0.7009	0.3228
Number of runs of positive and negative signs:						
Observed runs, m	3	3	3	3	16	11
Positive signs, p	71	42	53	16	28	61
Negative signs, n	39	67	56	93	21	48
Population mean, μ	51.35	52.63	55.46	28.30	25	54.72
Standard deviation, σ	4.774	4.920	5.192	2.579	3.391	5.121
t-value, $(\mu - m)/\sigma$	10.13**	10.09**	10.10**	9.812**	2.654*	8.538**
Parabolic regression of residuals with time:						
t-value of $\hat{\beta}_2$	-9.297**	-5.254**	2.227	5.489**	-0.3963	-8.250**
F for regression	43.21**	26.16**	7.935*	29.45**	0.3622	34.20**

* Significant at the 1% level.

** Significant at the 0.1% level.

TABLE 5.3.21

SUMMARY OF RESULTS FROM THE NONLINEAR MODELS M10, M13, M21 AND M22

STATISTIC	M10	M13	M21	M22
Number of observations	110	110	109	110
Estimates of the parameters:				
Variable parameter, u	561800	166300	1.233	+2.862E-17
Regression coefficient, b	-0.01837	4.010E-7	0.04695	-1.066E15
Error sum of squares	642600	885400	7.275E7	6.216E9
Standard error of the estimate	77.14	90.54	824.6	7587
Correlation between observed and calculatated values, r^2	0.9936**	0.9917**	0.9988**	0.9313**
Value of F for the hypothesis that $b_o = 0$	2.926	9.534*	0.5357	49.92**
Dependent variable	Δy	Δy	y	y

* Significant at the 1% level.

** Significant at the 0.1% level.

TABLE 5.3.22

STATISTICS OF THE RESIDUALS IN THE NONLINEAR MODELS M10, M13, M21 AND M22

STATISTIC	M10	M13	M21	M22
Number of residuals	110	110	109	110
Maximum deviation	217.6	289.0	2991	19759
Year of maximum	1956	1968	1968	1968
Range of deviations	421.7	458.0	5029	32560
Mean of residuals	-9.438	19.93	47.26	-3017
Median residual	-17.93	7.369	-36.08	-2051
Coefficient of skewness	0.6901	0.7270	0.2550	0.8766
t-value for skewness	2.995*	3.155*	1.101	3.804**
Coefficient of kurtosis	1.055	1.887	2.215	1.137
t-value for kurtosis	2.309	4.128**	4.826**	2.487*
Dependent variable	Δy	Δy	y	y

* Significant at the 1% level.

** Significant at the 0.1% level.

TABLE 5.3.23

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE NONLINEAR MODELS M10, M13, M21 AND M22

TEST STATISTICS	M10	M13	M21	M22
Durbin-Watson statistic	n.a.	n.a.	n.a.	n.a.
Number of runs of positive and negative signs:				
Observed runs, m	17	15	12	3
Positive signs, p	34	82	39	19
Negative signs, n	76	28	70	91
Population mean, μ	47.98	42.75	51.09	32.44
Standard deviation, σ	4.452	3.950	4.772	2.963
t-value, $(\mu - m)/\sigma$	6.959**	7.024**	8.192**	9.935**
Parabolic regression of the residuals with time:				
t-value of $\widehat{\beta_2}$	1.194	-2.790*	-1.031	15.98**
F for regression	0.9066	3.921	0.7008	132.6**

* Significant at the 1% level.

** Significant at the 0.1% level.

n.a. Test not applicable.

TABLE 5.3.24

SUMMARY OF RESULTS FROM THE NONLINEAR MODELS M9, M12, M16 AND M18

STATISTIC	M9	M12	M16	M18
Number of observations	110	110	110	110
Estimates of the parameters:				
Variable parameter, u	2.644	1999	325200	0.7767
Variable parameter, v	1.013	0.9297	0.2171	116100
Regression coefficient, b	1485	141900	0.1349	0.04235
Error sum of squares	1.023E9	1.973E7	5.953E5	6.522E5
Standard error of the estimate	3092	429.4	74.59	78.07
Correlation between observed and calculated values, r^2	0.9871**	0.9997**	0.9939**	0.9935**
Value of F for the hypothesis that $b_0 = 0$	30.58**	14.31**	0.04974	2.471
Dependent variable	y	y	Δy	Δy

** Significant at the 0.1% level.

TABLE 5.3.25

STATISTICS OF THE RESIDUALS IN THE NONLINEAR MODELS M9, M12, M16 AND M18

STATISTIC	M9	M12	M16	M18
Number of residuals	110	110	110	110
Maximum deviation	-9122	1150	191.5	180.6
Year of maximum	1968	1968	1956	1956
Range of deviations	14290	1896	382.2	359.2
Mean of residuals	-1131	-118.4	-1.209	-8.761
Median residual	-1416	-142.6	-1.459	-20.88
Coefficient of skewness	0.5270	0.8605	0.4707	0.7039
t-value for skewness	2.287	3.734**	2.043	3.055*
Coefficient of kurtosis	0.1495	0.6562	0.7752	0.6186
t-value for kurtosis	0.3272	1.436	1.696	1.353
Dependent variable	y	y	Δy	Δy

* Significant at the 1% level.

** Significant at the 0.1% level.

TABLE 5.3.26

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE NONLINEAR MODELS M9, M12, M16 AND M18

TEST STATISTICS	M9	M12	M16	M18
Durbin-Watson statistic	n.a.	n.a.	n.a.	n.a.
Number of runs of positive and negative signs:				
Observed runs, m	3	6	19	12
Positive signs, p	26	24	51	28
Negative signs, n	84	86	59	82
Population mean, μ	40.71	38.53	55.71	42.75
Standard deviation, σ	3.755	3.546	5.192	3.950
t-value, $(\mu - m)/\sigma$	10.04**	9.172**	7.070**	7.783**
Parabolic regression of the residuals with time:				
t-value of $\hat{\beta}_2$	5.167**	2.498	0.1453	0.7371
F for regression	20.33**	5.320	0.01923	0.6473

** Significant at the 0.1% level.

n.a. Test not applicable.

TABLE 5.3.27

SUMMARY OF RESULTS FROM THE NONLINEAR MODELS M15 AND M17

STATISTIC	M15	M17
Number of observations	110	110
Estimates of the parameters:		
Variable parameter, u	0.4362	0.8289
Variable parameter, v	0.01988	250900
Variable parameter, w	37.68	0.9999 (maximum)
Regression coefficient, b	526500	1.432E-6
Error sum of squares	7.458E6	6.045E5
Standard error of the estimate	265.3	75.52
Correlation between observed and calculated values, r^2	0.9999**	0.9939**
Value of F for the hypothesis that $b_o = 0$	6.694	0.7707
Dependent variable	y	Δy

** Significant at the 0.1% level.

TABLE 5.3.28

STATISTICS OF THE RESIDUALS IN THE NONLINEAR MODELS M15 AND M17

STATISTIC	M15	M17
Number of residuals	110	110
Maximum deviation	681.4	185.0
Year of maximum	1957	1956
Range of deviations	1154	366.8
Mean of residuals	51.69	-4.765
Median residual	57.86	-12.27
Coefficient of skewness	0.01594	0.6442
t-value for skewness	0.06917	2.795*
Coefficient of kurtosis	-0.2122	0.7192
t-value for kurtosis	-0.4644	1.574
Dependent variable	y	Δy

* Significant at the 1% level.

TABLE 5.3.29

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE NONLINEAR MODELS M15 AND M17

TEST STATISTICS	M15	M17
Durbin-Watson statistic	n.a.	n.a.
Number of runs of positive and negative signs:		
Observed runs, m	7	18
Positive signs, p	74	33
Negative signs, n	36	77
Population mean, μ	49.44	47.20
Standard deviation, σ	4.591	4.377
t-value, $(\mu - m)/\sigma$	9.243**	6.671**
Parabolic regression of the residuals with time:		
t-value of $\hat{\beta}_2$	-0.8584	0.3587
F for regression	1.914	0.1977

** Significant at the 0.1% level.

n.a. Test not applicable.

a function of time (as outlined below) the residuals were not random, and $t = 9.961$. (The results for model M3† are shown in Tables 5.3.9, 5.3.10 and 5.3.11.)

The two roots m_1 and m_2 in the second order differential equation used in model M3 turned out to be:

$$m_1 = 1.024,$$

and $$m_2 = 0.9953.$$

Referring to Equation 4.2.23, this implies that cumulative production is related to time as follows:

$y(t) = \beta_0 + \beta_1 m_1^{\ t} + \beta_2 m_2^{\ t}$ (5.3.1), with $\beta_0 = c/(1 + a + b)$ as in Equation 4.3.22. The regression of $y(t)$ on the right hand side of this equation using the above (but unrounded) values for m_1 and m_2, resulted in estimators of the coefficients as follows:

$y(t) = -1.386E5 + 1.040\, m_1^{\ t} + 1.327 m_2^{\ t}$ (5.3.2), with $t = 1, 2, 3, \ldots, 110$. If $t = 220$, then cumulative production would be 1,912,000 million barrels, with an annual production of 47120 million barrels. Projected over this span of time, model M3† thus implied some rather unrealistic results. (Production for the last two years of this series would be greater than the entire production over the first 110 years).

In models M5 and M8, there is a significant parabolic trend in the residuals with time, which suggests that any projection of production based on these models has a built-in bias with time.

It is interesting that models M4 and M6 each had the maximum residual for the year 1968. Thus, model M7 was the only multivariate model in which the residuals appeared to be random normal--except for the number of runs of positive and negative signs.

In many models, time (t) was used as a variable. In each case $t = 1, 2, 3, \ldots N$; for the years 1859 to 1968, $N = 110$. The notation used for the other variables has been outlined in Section One of this Chapter.

The intercorrelations between the variables in each multivariate model are shown in Tables 5.3.12, 5.3.13, 5.3.14, 5.3.15, 5.3.16, and 5.3.17. It would be difficult to pick variables which showed a higher relationship to production than these. Even if other (independent) variables were found which were valid indicators of the rate of production, there would be problems involved in projecting these variables over a long period of time, so as to anticipate the forecast for production.

The regression coefficients for six of the models (M1, M2, M11, M14, M19 and M20) could be estimated using linear regression, and the results are shown in Table 5.3.18. With the exception of model M19, the residuals in all these models had a significant parabolic trend with time, in all the models (except M19) the runs were definitely not randomly distributed. Also, the residuals were all significantly skewed and peaked, except in the case of model M2.

An estimate of the ultimate resource R was implied by some of these models. In model 11, R was calculated as 9150 million barrels. This figure is only a small fraction of the actual production up to 1968; the validity of the model is therefore highly questionable. Similarly, model M14 implied an estimate for R of 85.27 million barrels. In these cases, the estimate for R was derived from the intercept of the regression line.

For model M19, R was calculated to be less than 411300 million barrels (which would be recovered if percent recovery were projected to 100%). In model M20, R was calculated to be 130500 million barrels at the point where cumulative production reached cumulative discoveries. Thus, among all the models described above, M19 was the only one which provided an acceptable estimate for R.

In all the models discussed below, computer algorithms were required to estimate the parameters. It was assumed that if $y = o$, then $\Delta y = o$ in models M10, M13, M16, M18, and M17. Similarly, in model M21, it was assumed that if $q = o$, then $y = o$, and in M22 if $\Psi = o$, then cumulative production was assumed to be zero. Finally, in models M9, M12, and M15, it was assumed that if $t = o$, then $y = o$. In this latter case, a minor adjustment to the models originally proposed was necessitated.

In all the above models, regression through the origin could be employed, provided the assumptions made were true. In any case, the probability of the truth of the assumption that the intercept (b_o) is zero can be estimated by calculating the relative improvement in regression if the intercept was not assumed equal to zero.

The computer algorithm TEST (described in the Appendix) was used to minimize the residual sum of squares for the nonlinear models. In each of these models we have

$$z = z' + e \qquad (5.3.3)$$

where $z = z(t)$ = the observed values of the dependent variable,

$z' = z(t)$ = the values calculated to estimate z,

$e = e(t)$ = the error term, or residual, and $t = 1,2,3,\ldots N$.

Because the regression line was constrained to pass through the origin, we can write

$$z = bx + e, \qquad (5.3.4)$$

where $x = x(t)$ = the values calculated to estimate z, provided

$$b = \frac{\Sigma\, x(t)\, z(t)}{\Sigma (x(t))^2}, \qquad (5.3.5)$$

= the regression coefficient for regression through the origin.

In each of the models, the effect of excluding one regression coefficient can be tested by the ratio

$$F = (SSE' - SSE)/SSE/(N-2), \tag{5.3.6}$$

where SSE' = the residual sum of squares for error calculated using regression through the origin,

SSE = the error sum of squares corresponding to the regression equation,

$$z = b_o + bx + \epsilon, \tag{5.3.7}$$

and N = the number of observations. This ratio has an F distribution with 1 and N-2 degrees of freedom.[1]

The results for four nonlinear two parameter models (M10, M13, M21 and M22) are shown in Table 5.3.21. The hypothesis that the intercept is zero can be rejected here, except for model M21.

In order to calculate an F test for the significance of the regression through the origin, it is more convenient to use the notation z_i for $z(t)$ and x_i for $x(t)$, so that

$$zi = \beta xi + \epsilon_i \tag{5.3.8}$$

$$= bx_i + e_i \tag{5.3.9}$$

where β is the true regression coefficient, ϵ_i is the true error, and the least squares regression coefficient is $b = (\Sigma x_i z_i) / \Sigma x_i^2$.

Since $z_i = \beta x_i + \epsilon_i$,

then $\Sigma x_i z_i = \beta \Sigma x_i^2 + \Sigma \epsilon_i x_i$

$$\therefore \quad \frac{\Sigma x_i z_i}{\Sigma x_i^2} = \beta + \frac{\Sigma \epsilon_i x_i}{\Sigma x_i^2}$$

$$\therefore \quad b = \beta + e_b \tag{5.3.10}$$

where e_b = the error in b

$$= \frac{\Sigma \epsilon_i x_i}{\Sigma x_i^2}$$

$$= -(\beta - b) \tag{5.3.11}$$

Now SSE' = the minimum sum of squares for error,

$$= \Sigma e_i^2$$

$$= \Sigma (z_i - bx_i)^2$$

$$= \Sigma z_i^2 - 2b \Sigma x_i z_i + b \frac{\Sigma x_i z_i}{\Sigma x_i^2} \Sigma x_i^2$$

$$= \Sigma z_i^2 - b \Sigma x_i z_i ,$$

$$= \Sigma z_i^2 - (\Sigma x_i z_i)^2 / \Sigma x_i^2 \qquad (5.3.12)$$

Also, SSE'

$$= \Sigma (\epsilon_i - (b-\beta)x_i)^2$$

$$= \Sigma (\epsilon_i - e_b x_i)^2$$

$$= \Sigma \epsilon_i^2 + e_b^2 \Sigma x_i^2 - 2e_b \Sigma \epsilon_i x_i$$

$$= \Sigma \epsilon_i^2 - e_b^2 \Sigma x_i^2 \qquad (5.3.13)$$

The problem which remains is to describe the distribution of SSE'.

By hypothesis, the ϵ_i are normal, so that for all i,

$$E(\epsilon_i) = 0, \text{ and } V(\epsilon_i) = \sigma^2$$

Therefore, $E(e_b) = 0$

and

$$V(e_b) = \frac{\sigma^2 \Sigma x_i^2}{(\Sigma x_i^2)^2}$$

$$= \frac{\sigma^2}{\Sigma x_i^2} \qquad (5.3.14)$$

= the variance of b.

Now ϵ_i/σ has a standard normal distribution, so that $(\epsilon_i)^2/\sigma^2$ has a chi-squared distribution with one degree of freedom, and $\Sigma (\epsilon_i)^2 /\sigma^2$ has a chi-squared distribution with N degrees of freedom.[2]

Similarly $\frac{e_b \sqrt{\Sigma x_i^2}}{\sigma}$ has a standard normal distribution, and $\frac{e_b^2 \Sigma x_i^2}{\sigma^2}$ has a chi-squared distribution with one degree of freedom. Thus, using Equation 5.3.13 above, $\frac{SSE}{\sigma^2}$ has a chi-squared distribution with N-1 degrees of freedom.[3]

Thus we can write

$$E\ (SSE' / \sigma^2) = N-1, \tag{5.3.15}$$

$$E\ (SSE' / (N-1)) = \sigma^2, \tag{5.3.16}$$

and $$SSE' / (N-1) = s^2 \tag{5.3.17}$$

is an unbiased estimator of σ^2.

The estimate of the variance of b is $s_b^2 = s^2 / \Sigma x_i^2$, (5.3.18)

so that $$t = \frac{b - \beta}{s_b}, \tag{5.3.19}$$

has a t distribution with N-1 degrees of freedom.

If we hypothesize that $\beta = o$,

then $$t^2 = b^2/(s^2 / \Sigma x_i^2) \tag{5.3.20}$$

$$= (N-1)\, b^2\, \Sigma x_i^2 / SSE'$$

$$= F,$$

where F has an F distribution with 1 and N-1 degrees of freedom.[4]

The fact that the regression line is restricted to pass through the origin results in a small standard error in the regression coefficient b. An alternative test of whether the relationship $z = bx$ accounts for the variation in z significantly better than the assumption that

$$z = b_o x_o,$$

where $x_o = 1$,

so that $b_o = \bar{z}$, the mean of z, is given by the equation

$$F = \frac{(\Sigma xz - \Sigma z\, (\Sigma x^2 / N)^{1/2})^2}{2 s_o^2\, (\Sigma x^2 - \Sigma x\, (\Sigma x^2 / N)^{1/2})^2} \tag{5.3.21}$$

where s_o^2 is the residual mean square of z, with N-2 degrees of freedom. This F test has 1 and N-2 degrees of freedom, and may yield more appropriate values for F than the simple test of the departure of b from zero-especially if the data consist of a cluster of points at some distance from the origin.[5]

In all the models, the correlation between the right-hand side and the left-hand side of the equation used was highly significant. The F test results were not included in the tables of results; it is obvious from the correlation coefficient calculated and shown in the results for each model that a statistically significant relationship existed.

Even in model M22 (where the hypothesis that cumulative production would be asymptotic to some upper value could be rejected because of the estimates obtained for the parameters) there was a high coefficient of correlation between the dependent and independent (calculated) values. Thus, sequential data covering over one hundred years implied such high levels of significance, it was difficult to choose among the models using the coefficient of correlation alone.

In model M22, values such as $\delta = 0$, 1, and 5 were tried in the equation $y(t+\delta) = R(1-e^{c\Psi})$. The smallest sum of squares for error was obtained for $\delta = 0$. The model did not converge to a positive value for R in any case, and so the model could be rejected. This simply highlights the problem of finding an independent variable which could be employed to impute an upper limit for cumulative production. Cumulative production of crude oil in the United States has been increasing at a rapid rate, greater than could be implied by model M22 using the number of wells drilled as an independent variable.

With the above exception, the three and four parameter models naturally tended to yield more desirable results than the models where fewer parameters were estimated. This was evidenced in the sum of squares for error, the number of runs of positive or negative signs, and the parabolic regression of the residuals with time.

However, in models M9 and M12, the hypothesis that the regression line passed through the origin could be rejected at the 0.1% level. (In model M15, the rejection of this hypothesis could be made at the 2 1/2% level.) The results

from the nonlinear models which had more than two parameters to be estimated simultaneously are shown in Tables 5.3.24, 5.3.25, 5.3.26, 5.3.27, 5.3.28, and 5.3.29. Note that the mean of the residuals was not zero in the nonlinear models because the regression line passed through the origin. (In ordinary linear regression, the mean of the residuals would be zero by definition.)

These models in which only two or three parameters were to be estimated simultaneously presented no special problems(so far as rapid convergence to definite values of the parameters were concerned). However, models M15 and M17 (where four parameters including the regression coefficient were estimated simultaneously) converged very slowly; as soon as the accuracy of the computer algorithm was increased, the parameters tended to drift to quite different values.

For this reason, computer algorithms used are not recommended for models such as M15 and M17. Furthermore in model M17, the values for R and n in the equation

$$\Delta y = ky^m (R-y)^n \tag{4.2.34}$$

tended to increase beyond the capability of the IBM360/67 APL system. Thus, it was decided to restrict n to an upper limit of 0.9999. This was in keeping with the nature of the original model proposed, where

$0 < m < 1$, and $0 < n < 1$,

so that $0 < t < s$, as in equations 4.2.35 and 4.2.35. That is, if

$$\overset{\circ}{y} = \frac{dy}{dt} = ky^m (R-y)^n \tag{4.2.34}$$

then

$$y^{-m} (R-y)^{-n}\, dy = k\,dt \tag{5.3.22}$$

so that

$$\int_0^R y^{-m} (R-y)^{-n}\, dy = \int_0^s k\,dt \tag{5.3.23}$$

if the resource R is produced during the time span $0 < t < s$. Now, if we put $x = y/R$, then

$$R^{m+n-1} \int_0^1 x \quad x^{-m}(1-x)^{-n}\, dx = \int_0^s k dt \qquad (5.3.24)$$

$$\therefore \qquad R^{m+n-1} \quad B\,(\,(1-m),\,(1-n)\,) = ks \qquad (5.3.25)$$

It can be seen that if s becomes very large, then R becomes very large, since $B\,(\,(1-m),\,(1-n)\,)$ is finite, provided $m < 1$ and $n < 1$. If m or n are greater than (or equal to) unity, then $B(\,(1-m),\,(1-n)\,) = \Gamma(1-m)\,\Gamma(1-n)/\Gamma(2-m-n)$ may become unbounded in size, or negative, and the equation breaks down. For $n=23$, $\Gamma(1-n) = -\Gamma(22)$, and this is infinite.

Using values for every five years (in order to speed calculations) the last optimal values for m, n and R calculated by the computer algorithm were $m = 0.79$, $n = 23$, and $R = 5.5 \times 10^6$ million barrels, with both n and R increasing. Thus, any further attempts to obtain a least squares fit with m, n and R allowed to vary arbitrarily were discontinued in model M17, and n was set at 0.9999 for all other calculations.

In any of the models where the test of the Durbin-Watson statistic could be applied, there was a positive serial correlation at the 5% level of significance. For example, at the 5% level of significance, with one independent variable and a sample size of 100, the lower limit for the Durbin-Watson test statistic would be 1.65, and any lower value would confirm the hypothesis of positive serial correlation. In each model studied herein, this statistic was never greater than unity.

REFERENCES:

1. K.W. Smillie, An Introduction to Regression and Correlation, The Ryerson Press, Toronto, 1966, p.27.

2. E.S. Keeping, Introduction to Statistical Inference, D. Van Nostrand Company, Inc., Toronto, 1962, p.87.

3. Ibid, p.88.

4. E. S. Keeping, Personal Communication, July, 1970.

5. E. J. Williams, Regression Analysis, John Wiley & Sons, Inc., New York, 1959, pp. 77, 78.

CHAPTER VI

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

This thesis is concerned with the long term pattern of utilization of crude oil. In order to limit the scope of the research to a level appropriate for a study of this kind, only one region was investigated, namely the continental United States, excluding Alaska. While there was reason to believe that the oil industry (which is international in nature, and dominated by eight large firms) could best be described using data for a more global area, it was hoped that the present study would at least offer a basis for such a project.

Within the United States, the rapid rate of utilization of oil has captured the attention of a number of investigators. Various models of the time pattern of oil discovery and of production have been proposed. The main criticism of the models based on discoveries has been directed toward the unreliability of such data for proved reserves, and for oil in place. On the other hand the major argument against using just the statistics on actual production has been that these models are arbitrary, and difficult to reconcile with geological information on proved reserves and discoveries of oil in place.

In the present study, a number of models were proposed to project oil production. It was supposed that a comparison of the results obtained would offer some means for choosing among the models hypothesized. Any residual pattern in the error terms could also receive particular attention.

Even though the models were merely an abstraction of reality, whenever possible, the models hypothesized were devised to meet certain boundary conditions which were believed to apply at the beginning (and at the end) of the cycle of discovery and utilization of oil. In a few cases, the probability that such hypotheses were true (as opposed to the alternative hypothesis that they were not true for the model) could be estimated.

As outlined in Table 5.1.1, a total of twenty-two models were proposed to describe the pattern of oil production for the United States during the period 1859 to 1968. Also, as described in Section One of the previous chapter, data were not available for some models for the entire period.

Twelve of the models were linear, and their coefficients could be estimated using ordinary linear regression. Ten of the models were nonlinear, and the computer algorithms described in Appendix H were devised to estimate the various unknown parameters or coefficients. By minimizing the sum of squares for error with the data used in a model, the computer algorithms could calculate estimates of one, two or three parameters, and also the regression coefficient, assuming regression through the origin. All nonlinear models were hypothesized in such a way that regression through the origin was necessarily used.

In each of the twenty-two models, it was apparent that the residuals were not random so far as the number of runs of positive and negative signs were concerned. This implies that either none of the models adequately describes the long term utilization of oil in the United States, or that there are residual short term cycles of highs and lows in production to be accounted for.

Six of the twelve linear models were multivariate in nature; none of the multivariate models (M3, M4, M5, M6, M7 and M8) gave any estimate of the ultimate resource, R. Of the remaining six univariate models (M1, M2, M11, M14, M19 and M20) only model M19

$$y = p.r./m - b/m \qquad (4.5.1)$$

imputed an estimate for R which was within the range of 193 to 2319 billion barrels, which was set as a rough criterion in Section Two of Chapter Five.

Except for model M19, the residuals in the univariate models had a significant parabolic regression with time. These results, and the results for the nonlinear models, have been summarized in Tables 6.1.1, 6.1.2, and 6.1.3.

continued on p.143

TABLE 6.1.1

SUMMARY OF SOME RESULTS FROM MODELS OF TRANSFORMED VALUES ON CUMULATIVE PRODUCTION

Model	Assumed Relationship (and Comments)	Standard Error of the Estimate	N	Estimate of R
M7	$\ln y = \beta_0 + \beta_1 \ln q + \beta_2 \ln w + \beta_3 \ln f$	0.02285	35	None
	(The parabolic trend of residuals with time is not significant; $t = -0.05270$.)			
M8	$\ln y = \beta_0 + \beta_1 \ln q + \beta_2 \ln \Psi$	0.1776	109	None
	(Parabolic trend of residuals with time: $t = 2.109$; $\alpha < 5\%$.)			
M1	$\ln y = \ln k + ct$	1.072	110	∞
	(Parabolic trend of residuals with time: $t = -9.297$; $\alpha < 0.1\%$.)			
M11	$\ln(y(t+1)) = (1-d)(\ln R) + d \ln(y(t))$	0.4488	109	9150
	(Parabolic trend of residuals with time: $t = 2.227$; $\alpha < 5\%$.)			
M20	$\ln y = \ln b + c \ln q$	0.3936	109	130500
	(Parabolic trend of residuals with time: $t = 73.47$; $\alpha < 0.1\%$.)			
M14	$1/y(t+1) = (1-b)/R + b/y(t)$	0.04316	109	85.27
	(Parabolic trend of residuals with time: $t = 64.24$; $\alpha < 0.1\%$.)			

TABLE 6.1.2

SUMMARY OF SOME RESULTS FROM MODELS OF ANNUAL PRODUCTION

Model	Assumed Relationship (and Comments)	Standard Error of the Estimate	N	Estimate of R
M5	$\Delta y(t) = \beta_0 + \beta_1 D(t) + \beta_2 G(t) + \beta_3 H(t) + \beta_4 N(t) + \beta_5 r(t)$ (Parabolic trend of residuals with time: $t = -2.339$; $\alpha < 5\%$.)	43.23	49	None
M6	$\Delta y = \beta_0 + \beta_1 (D(t) - \Delta y(t)) + \beta_2 (H(t) - \Delta y(t)) + \beta_3 (N(t) - \Delta y(t)) + \beta_4 r(t)$	107.1	49	None
M10	$\Delta y = -ay \ln (y/R)$	77.14	110	561800
M13	$\Delta y = cy(R-y)$ (Model rejected because $b_0 \neq 0$, $\alpha < 1\%$. Parabolic trend of residuals with time: $t = 2.790$; $\alpha < 1\%$.)	90.54	110	166300
M16	$\Delta y = ky(1-(y/R)^{\beta})$	74.59	110	325200
M18	$\Delta y = ky^{m} (R-y)^{1-m}$	78.07	110	116100
M17	$\Delta y = ky^{m} (R-y)^{0.9999}$	75.52	110	250900

TABLE 6.1.3

SUMMARY OF SOME RESULTS FROM MODELS OF CUMULATIVE PRODUCTION

Model	Assumed Relationship (and Comments)	Standard Error of the Estimate	N	Estimate of R
M3	$y(t) = c - ay(t-1) - by(t-2)$	62.29	108	See $M3^{\dagger}$
$M3^{\dagger}$	$y(t) = \beta_0 + \beta_1 m_1^{\ t} + \beta_2 m_2^{\ t}$	1735	110	∞
M4	$y(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4$	461.2	110	∞
M2	$y(t) = b + ay(t-1)$ (Parabolic trend of residuals with time: $t = -5.254$; $\alpha < 0.1\%$.)	219.8	109	∞
M19	$y = p.r./m - b/m$	2200	49	<411300
M21	$y = b\,q^{c}$	824.6	109	492400
M22	$y = R(1 - e^{c\Psi})$ (Model rejected because $b_o \neq 0$; $\alpha < 0.1\%$. Parabolic trend of residuals with time: $5 = 15.98$; $\alpha < 0.1\%$.)	7587	110	-1.066E15
M9	$y = R\,c^{d^t}$ (Model rejected because $b_o \neq 0$; $\alpha < 0.1\%$. Parabolic trend of residuals with time: $t = 5.167$; $\alpha < 0.1\%$.)	3092	110	1485
M12	$y = R/(1 + Ab^{t})$ (Model rejected because $b_o \neq 0$; $\alpha < 0.1\%$. Parabolic trend of residuals with time: $t = 2.498$; $\alpha < 2\%$.)	429.4	110	141900
M15	$y = R/(1 + Ae^{-at})^{\theta}$ (Model rejected because $b_o \neq 0$; $\alpha < 2.5\%$.)	265.3	110	526500

Except for model M7, all models in which some transformation was made of the dependent variable had residuals with a significant parabolic trend with time. However, model M7 was a multivariate model, and did not yield any estimate of R.

Of the ten nonlinear models, four had one parameter to be estimated (in models M10, M13, M21 and M22), four had two parameters to be estimated (in models M9, M12, M16 and M18), and two had three parameters to be estimated M15 and M17).

Regression through the origin was assumed for all the nonlinear models; in models M13 (the logistic curve for annual production), M22 (cumulative production as a function of the number of wells drilled), M9 (the Gompertz curve for annual production), M12 (the logistic curve for cumulative production), and M15 (the generalized logistic curve for cumulative production), this hypothesis (and thus the validity of the model) could be rejected at the 1%, 0.1%, 0.1% 2% and 2.5% levels respectively.

This left only one nonlinear model for cumulative production (M21), and four models for annual production (M10, M16, M18 and M17) which could not be rejected because of

i) a parabolic trend of the residuals with time,

or ii) the regression coefficient b_o was not equal to zero.

As might have been anticipated, these nonlinear models (with one or more parameters to be estimated) had a much smaller standard error of the estimate. Models M16 and M17 had the lowest standard error. In model M16, the sum of squares for error was 595300 for 110 observations, versus a sum of squares for error of 604500 in model M17 for 110 observations. Using an F ratio to discriminate between these two models,

$$F = (604500 - 595300) / 595300 / 107 \doteq 1.7$$

This means the probability is greater than 0.75 that model M16 offers a significant improvement over its nearest rival (M17), in terms of least squares. As can be seen in Table 6.1.2, model M16 has a much smaller standard error of the estimate than any of the unrejected models for annual production. For these reasons, model M16 (the generalized logistic curve) was a natural candidate to project annual production. The resource base R implied by model M16 was 325.2 billion barrels of crude oil, using all the data from 1859 to 1968.

A few additional tests were made using model M16. Using data for the even years only from 1859 to 1968, model M16 gave an estimate of R = 331.7 billion barrels. Using odd years only, R was estimated to be 316.5 billion barrels. The first 55 years, from 1859 to 1913 implied an estimate of R = 4.062 billion barrels; the last 55 years from 1914 to 1968 gave an estimate for R of 353.9 billion barrels. (The low estimate obtained using the first 55 years can be understood from the fact that the entire cumulative production up to 1913 was only 3.070 billion barrels, or approximately the same as the annual production for 1968.)

The residuals obtained when model M16 was used with data from 1914 to 1968 did not have a significantly low number of runs for either the odd years or the even years. This suggested that a short term cycle of highs and lows in production existed during the period, and that model M16 could not be rejected on the basis of any of the tests imposed upon it, as shown in Appendix M.

However, when data for cumulative discoveries of crude oil were used in model M16 instead of cumulative production, the estimate obtained for R was 162.7 billion barrels. This figure was less than the range for R already established, and so model M16 could not be used with cumulative discovery data to calculate Δq and an implicit estimate of R.

As outlined above, models M19 and M21 were the only models for cumulative production not rejected because of a parabolic trend of the residuals with time, or because $b_0 \neq 0$, and which gave an acceptable estimate for R. Model M19 implies a resource base of 411.3 billion barrels at 100% recovery. At 80% ultimate percent recovery, model M19 provides about the same estimate for R as does M16. Both models M19 and M21 for cumulative production had a high standard error of the estimate, which reflects the relatively high variation inherent in data on discoveries of oil.

Model M21 was a very coarse model, and was included only to arrive at some initial functional relationship between cumulative production and cumulative discoveries. Apart from the observed runs of positive and negative signs in the residuals, and a significantly high coefficient of kurtosis, model M21 proved to be a reliable model. Perhaps M21 would have been more appropriately proposed in the form

$$y = \beta_0 + \beta_1 q^c \qquad (6.1.1)$$

in order to allow for the fact that q could be greater than zero when y was zero. However, the improvement in the sum of squares for error was not appreciable, and the estimate of the resource base R relatively unchanged. R = 492.4 billion barrels in model M21, and R = 478.7 billion barrels in model M21*. (The ten nonlinear starred models are discussed below.)

In order to determine the effect of restricting the regression line to the origin, all ten nonlinear models were reformulated in such a way as to allow β_0 to assume any value in the equations

$$y = \beta_0 + \beta_1 g(x), \qquad (6.1.2)$$

or $$\Delta y = \beta_0 + \beta_1 h(x), \qquad (6.1.3)$$

where g(x) was previously defined in the nonlinear models by either

$$y = \beta g(x), \qquad (6.1.4)$$

or $$\Delta y = \beta h(x), \qquad (6.1.5)$$

because regression through the origin had been stipulated. Models M10, M13, M21, M22, M9, M12, M16, M18, M15 and M17 were thus restated as M10*,

M13*, M21*, M22*, M9*, M12*, M16*, M18*, M15* and M17*. For example, Equation 6.1.1 became model M21* with the introduction of the regression coefficient β_0.

With β_0 arbitrary, the determination of the parameters was essentially the same as for $\beta_0 = 0$, except that the sum of squares for error, SSE, was given by

$$SSE = \sum_{t=1}^{N} (z_t - b_o - b_1\, g(x_t))^2 \qquad (6.1.6)$$

where z_t = the observed dependent variable (either $\Delta y(t)$ or $y(t)$),

x_t = the observed independent variable used in a model,

$g(x_t)$ = the value of the observed independent as transformed by the particular function used in a model,

b_1 = the regression coefficient which was an estimate of β_1,

and b_o = the regression coefficient which was an estimate of β_0.

Since $b_o = \bar{z} - b_1 \bar{g}$ (6.1.7)

where $\bar{z}$ = the mean of z_t,

and $\bar{g}$ = the mean of $g(x_t)$,

b_o can be derived from b_1. From the normal equations for linear regression with β_1, b_1 would be given by

$$b_1 = \frac{\Sigma (z_t - \bar{z})\,(g(x_t) - \bar{g})}{\Sigma (g(x_t) - \bar{g})^2} \qquad (6.1.8)$$

Thus, in the nonlinear equations, the unknown parameters could be estimated in the same manner as for the case of regression through the origin. The same computer algorithms were used, except that the sum of squares for error was calculated as shown in Equation 6.1.6. For regression through the origin the sum of squares for error was calculated using Equation 5.3.12 with

$$b = \frac{\Sigma z_t\, g(x_t)}{(\Sigma g(x_t))^2}, \qquad (5.3.5)$$

where b was an estimate for the single regression coefficient β, as outlined in the previous chapter. The computer algorithms used for allmodels are shown in Appendix H.

The supplementary results obtained for the ten starred models are given in Appendix K, and are summarized in Table 6.1.4.

When the nonlinear models were formulated so as to use a regression line not necessarily through the origin, estimates of the parameters in model M9* (the Gompertz curve) were quite different. In the other models this effect was not quite so marked.

The residuals in model M16 had a slightly smaller standard error than the residuals for model M16*, which was probably due to the fact that model M16* had one extra degree of freedom for error because regression through the origin was not specified for the starred models.

Among these nonlinear models for cumulative production, M15* (the generalized logistic) had by far the lowest standard error of the estimate. Model M15* implied a resource base estimate of R = 490.6 billion barrels. Using data for the 55 years 1859 to 1913, the computer algorithm estimated a value of R = 6.7E31 million barrels, with the algorithm not converged, and R still increasing. Using the last 55 years from 1914 to 1968, M15* gave an estimate for R of 477.2 billion barrels.

This was an interesting result, and model M15* may even represent an improvement over model M16, which was favoured earlier.

However, lacking any way of discriminating between models M16 and M15*, considering the problem of estimating the three parameters simultaneously in M15*, and finally because M16 met all the boundary conditions set out earlier, (which M15 and M15* did not) the reasonable candidate among all the models studied would have to be model M16.

Model M16 gave an estimate for R which was consistent with model M19. Model M16 allows for the fact that cumulative production is calculated simply by adding annual production for all the years. And finally, at the point in time when annual production was zero, so was cumulative production. (It would not be necessary to use data for all the earlier years in model M16, and the exact time when production was zero is independent of the cycle of increase and eventual decline in production which would of necessity occur during the entire process of utilization of the resource base.)

TABLE 6.1.4

COMPARISON OF SOME ESTIMATES OBTAINED FOR THE NONLINEAR MODELS, OBTAINED BY USING REGRESSION THROUGH THE ORIGIN, AND REGRESSION NOT RESTRICTED TO THE ORIGIN

Model	Dependent Variable	Regression Through the Origin		Model	Regression Not Restricted to the Origin	
		Sum of Squares for Error	R		Sum of Squares for Error	R
M10	Δy	642600	561800	M10*	620800	522600
M13	Δy	885400	166300	M13*	806000	170700
M21	y	7.275E7	492400	M21*	7.218E7	478700
M22	y	6.216E9	-1.066E15	M22*	6.843E8	-20770
M9	y	1.023E9	1485	M 9*	6.902E6	546400
M12	y	1.973E7	141900	M12*	1.535E7	147300
M16	Δy	595300	325200	M16*	594800	333000
M18	Δy	652200	116100	M18*	627500	123300
M15	y	7.458E6	526500	M15*	6.577E6	490600
M17	Δy	604500	250900	M17*	597700	249000

A comparison of the results obtained in models M16 and M15* for the first 55 years (from 1859 to 1913) versus the results obtained during the second 55 years (from 1914 to 1968) is instructive. It appears that M16 may give a conservative, or low, estimate for R, because the estimate for R increases for the more recent data. On the other hand, model M15* produces estimates which decrease in value when the more recent data is used. Based on these two models, the resource base for the continental United States would be within the range of 353.9 to 477.2 billion barrels. Thus, depending upon one's degree of optimism, a figure could be selected within this range as one of the parameters for model M16. For example, if a figure of 400 billion barrels were used, only one parameter in M16 would have to be estimated, because from Equation 4.2.33 we could write

$$\Delta y = by(1 - (y/400000)^p), \qquad (6.1.9)$$

with p a parameter to be determined in this nonlinear equation, and b the regression coefficient estimated on the basis of the value determined for p.

The value of p and R estimated by model M16 was $p = 0.2171$, and $R = 325200$. Using data from 1934 to 1968, if R was set at a value of 400000, p was estimated to be 0.1412; with these values, Equation 6.1.9 can be written as

$$\Delta y \doteq (by/6.18)(6.18 - y^{0.1412}), \qquad (6.1.10)$$

because $6.18 \doteq 400000^{0.1412}$.

This equation gives some notion of the relative effect of the term $(R^p - y^p)$ describing the decline in Δy. The effect of this term would be small at first, when y was small, but would become pronounced as y increased. As y increases, the term rapidly tends to zero, because of the value for p.

Another way of illustrating the relative effect of the term constributing to decline in annual production is offered by model M17, where

$$\Delta y = ky^m (R - y)^n \qquad (4.2.34)$$

In this model, the value estimated for m was 0.8289, with the upper limit for n set at 0.9999.

The value estimated for R was 250.9 billion barrels of oil. Now Equation 4.2.34 can be restated as

$$\ell n\, \Delta y = \ell n k + m \ell n y + n \ell n (R-y) \qquad (6.1.11)$$

In this expression m and n may be considered as weights in a multivariate equation for $\ell n \Delta y$, which is linearly related to $\ell n y$ and $\ell n(R-y)$. Because n was found to be greater than m, the (R-y) term contributing to the decline in Δy would be given greater weight than the term for cumulative production.

The foregoing results suggest that annual production must now be reaching a maximum. The problem, however, is to find some variable which would serve to manifest the convergence of cumulative production to some upper limit, R.

In model M22, cumulative production did not converge with the cumulative number of wells drilled. The following statistics may indicate why this was so. The period 1934 to 1968 accounted for over 80% of the cumulative production, over 75% of the cumulative discoveries, and over 50% of the cumulative discoveries of oil in place. This same period accounted for about 60% of all wells drilled, but over 70% of the cumulative footage of wells, - all within the last 35 years of the 110 year history of the oil industry.

An examination of the data for these variables indicates that the annual number and annual footage of wells drilled reached a maximum in 1956. On the other hand, annual production has usually increased every year - right up to 1968. As discussed in Chapter III, annual discoveries of oil in place (classified by discovery year) are revised and updated for several years as more information becomes available. However, the relative appreciation inherent to these estimates soon tapers off, and it is therefore reasonable to assume that annual discoveries of oil in place were at their height sometime before 1951. The data for annual discoveries of proved reserves, Δq, while not subject to appreciation (because these figures are simply new reserves proved from all fields during a given year) are quite variable from year to year, but a maximum appears to have been reached during the period 1948 to 1951.

As an example of the generalization of the techniques developed in this study, consider a modified version of model M22 to project cumulative discoveries of oil in place, cumulative production, and cumulative discoveries. With some

benefit of hindsight, we can compare results from the following six cases in which cumulative footage of wells is used as the independent variable, rather than cumulative number of wells as in model M22.

Case I: $u(t) = \beta\,(1 - e^{-cf(t)})$, $t = 1,2,3,\ldots,35.$ (6.1.12)

Case II: $u(t) = \beta_0 + \beta_1\, e^{-cf(t)}$, $t = 1,2,3,\ldots,35.$ (6.1.13)

Case III: $u(t) = \beta_0 + \beta_1\, e^{-cf(t)}$, $t = 1,2,3,\ldots,20.$ (6.1.13)

Case IV: $u(t) = \beta_0 + \beta_1\, e^{-cf(t)}$, $t = 1,2,3,\ldots,15.$ (6.1.13)

Case V: $y(t) = \beta_0 + \beta_1\, e^{-cf(t)}$, $t = 1,2,3\ldots,35.$ (6.1.14)

Case VI: $q(t) = \beta_0 + \beta_1\, e^{-cf(t)}$, $t = 1,2,3,\ldots,35.$ (6.1.15)

The following notation was repeated for convenience:

$u = u(t)$ = cumulative discoveries of oil in place at the end of year t, classified by year of discovery.

$y = y(t)$ = cumulative production of crude oil at the end of year t.

$q = q(t)$ = cumulative discoveries of crude oil at the end of year t.

$f = f(t)$ = cumulative footage of all wells drilled at the end of year t. (Data on footage of wells drilled was available only for the period 1934 to 1968.)

$t = 1,2,3,\ldots 35$ represented the years from 1934 to 1968 inclusive in Case I, Case II, Case V and Case VI.

$t = 1,2,3,\ldots,20$ represented the years from 1934 to 1953 inclusive in Case III.

$t = 1,2,3,\ldots,15$ represented the years from 1954 to 1968 inclusive in Case IV.

β, β_0, β_1 were the unknown population regression coefficients estimated by b, b_0, and b_1 respectively.

c was a parameter to be determined by the method of least squares, using the computer algorithms TEST1 or TEST2 described in Appendix H.

δ was a correction factor for the initial cumulative footage at the end of 1933. (As described in Section One of Chapter V, cumulative footage up to 1933 was estimated at 2,156,663 thousand feet, using a figure of 2500 feet per well drilled.) δ can be calculated from c, b_0 and b_1, as indicated below.

All the above models are fashioned on a general model of the form

$$u = U\left(1-e^{-c(f+\delta)}\right), \qquad (6.1.16)$$

where U was the upper limit to which u would converge as f became very large. This equation can be restated as

$$u = \beta_0 + \beta_1 e^{-cf} \qquad (6.1.13)$$

where $\beta_0 = U$,

and $\delta = -(\ln \beta_1/\beta_0) \div c$

The two parameters c and δ, and the regression coefficient U in Equation 6.1.16 could be estimated using the computer algorithm TEST2. Similarly, the computer algorithm TEST1 could be used for Equation 6.1.13. However, it was simpler to estimate only one parameter c (and the two regression coefficients) in the latter equation. The same results were obtained in either case, provided δ was allowed to be arbitrary. Therefore, the parameter c, and the various coefficients (which were dependent upon c) were estimated using the algorithm TEST1 for Case II, Case III, Case IV, Case V and Case VI. Some results from all six cases have been summarized in Table 6.1.5. More detailed results are contained in Appendix L.

The results for Case III and Case IV indicate that these models for cumulative footage to predict cumulative discovereis of oil in place cannot be rejected, except for the fact that the residuals were serially correlated at the 5% level of significance. However, the Durbin-Watson test statistic indicated a positive serial correlation among the residuals in every model where the test could be applied. It was assumed that some short-term pattern of highs and/or lows existed in the data for y, q, u and f, and inasmuch as the Durbin-Watson test did not serve to distinguish the better models in any case, this test was rejected as

TABLE 6.1.5

SOME RESULTS FROM PROJECTIONS OF CUMULATIVE DISCOVERIES OF OIL IN PLACE, CUMULATIVE PRODUCTION AND CUMULATIVE DISCOVERIES, FROM CUMULATIVE FOOTAGE OF WELLS DRILLED

Case	Estimates of the regression coefficients and the correction factor, δ			
	b	b_0	b_1	δ
I	442.7			
	(This model could be rejected because of a parabolic regression of the residuals with time significant at the 0.1% level.)			
II		393.0	-807.2	-1123
	(This model could be rejected only because the number of runs of signs of the residuals was not random at the 1% level of significance.)			
III		388.0	-851.2	-1164
IV		397.7	-594.0	-732.9
V		375.5	-391.9	-1044
	(This model could be rejected only because the number of runs of signs of the residuals was not random at the 0.1% level of significance.)			
VI		161.2	-205.5	-1199
	(This model could be rejected only because the number of runs of signs of the residuals was not random at the 1% level of significance.)			

Note: All the above figures should be multiplied by a factor of 1000 to make them consistent with the data used. Thus, the regression coefficients given above are in units of billions of barrels, and the correction factor δ shown is in units of millions of feet.

a criterion for choosing among the long term models.

In Case II, Case III and Case IV the estimates of ultimate cumulative discoveries of oil in place (U) was 393.0, 388.0 and 397.7 billion barrels of oil. The ultimate recoverable resource of oil (R) would be less than the ultimate discovery of oil in place. However, the annual discoveries of oil in place should have been appreciated to their estimated ultimate value, as discussed in Chapter III for annual discoveries of proved reserves. Lacking any data to do this, and because estimates of annual discoveries of oil in place appreciate to approximately their ultimate value in about five years or so, then on the basis of the results from Case III and IV, the upper limit for R could be set at 400 billion barrels oil. As can be seen from Table 6.1.4, this criterion for R would result in the rejection of the nonlinear models M10, M10*, M21, M21*, M9*, M15 and M15*, apart from any of the other tests imposed on these models. Since all these models had either been pre-empted by other models with a lower sum of squares for error, or had been rejected on the basis of some other criterion, and had no special redeeming features in any case, their rejection forthwith presented no problem. Models M16, M16*, M17 and M19, and Case V were therefore the only five of all the models proposed which implied an acceptable value for ultimate cumulative productive.

While the models for Case V and Case VI could be rejected because of the small number of runs of signs among the residuals, the results obtained as an estimate of the upper bound of the recoverable resource R was instructive. In Case V, using cumulative production, the upper limit for y would be about 380 billion barrels if f became very large. The standard error of the estimate was much less than for model M19 - the only other model for cumulative production not rejected vecause of the value estimated for R. However, in Case VI, the upper limit for q would be about 160 billion barrels if f became very large, and so this model could be rejected on the basis of this value estimated for R.

The relationship which had been assumed in this study between y and q is given by Equations 4.5.3 and 4.5.4 in models M20 and M21, respectively. While model M21 appeared to be among the better of all the models proposed, the relationship between y and q apparently may not be expressed just as a simple power function. Model M20 was rejected because of a parabolic regression of the residuals with time. Also, if we assumed that cumulative production followed a similar pattern to cumulative discoveries in time (or footage of wells drilled), and therefore tried cumulative discovery data in a model used for production, then a value for R was obtained which was less than the lower limit for R which had been established. The nature of the relationship between y and q was thus not revealed by the models used in this study.

The above results point out the need to establish a more functionally consistent set of models linking the main variables y, q, u and f.

This suggests the necessity for a different set of relationships between the main variables y, q, u and f used in this study. In Figure 2.3.1, the interrelationship between some of the factors in the discovery and utilization of oil is implied. This relationship can be restated in a simpler form, as follows:

$$\begin{array}{ccc} f & \longrightarrow & u \\ \uparrow & & \downarrow \\ y & \longleftarrow & q \end{array} \qquad (6.1.17)$$

where, for example, $f \rightarrow u$ means the amount of oil in place, u, is implied by cumulative footage of wells drilled, f. The model for Case III and IV describes this relationship reasonably well.

To model the relationship between cumulative production and cumulative footage of wells drilled, Case V might be restated as

$$e^{-cf} = \frac{\beta_0 - y}{-\beta_1} \quad , \qquad (6.1.18)$$

or $$f = \frac{1}{c}\,\ell n\,(-\beta_1) + \frac{1}{c}\,\ell n\,(\beta_0 - y) \quad ,$$

$$= F + \frac{1}{c}\,\ell n\,(\beta_0 - y) \quad , \qquad (6.1.19)$$

where F = the upper limit of the cumulative footage of wells drilled, and which would be reached if $y = R = \beta_0$ (less one barrel, say), the upper limit of cumulative production.

In Case III we had

$$u = U + \beta_1 e^{-af} \qquad (6.1.20)$$

which can be restated as

$$f = F + \frac{1}{a} \ln(U-u) \qquad (6.1.21)$$

Combining this equation with Equation 6.1.19 results in

$$a \ln(R-y) = c \ln(U-u) \qquad (6.1.22)$$

or

$$y = R-(U-u)^{c/a} \qquad (6.1.23)$$

Thus if the models for Case III and Case V are correct, then model M7 might be improved by incorporating the relationship implied by Equations 6.1.23 and 6.1.19 into the model. It would be of interest to test this hypothesis. Also, because model M20 was rejected, model M7 could be improved further by substituting a more adequate relationship between y and q.

For this purpose, it would probably be worthwhile to first study the nature of the relationship between f and q. The number of barrels of proved reserves Δq discovered by Δf footage of wells, (or conversely, the footage of wells drilled as a result of Δq new discoveries) should be examined.

More than likely, it would be worthwhile to separate total footage into wildcat footage and development footage in such an investigation. There probably should be close relationship between wildcat footage and discoveries of oil in place, whereas development footage should be related to production, rather than to discoveries of oil in place. Also, if discoveries of new proved reserves, Δq, could be separated into additions to reserves from new fields, and additions to reserves as a result of development of fields already discovered, then improved models might be formulated, and proven superior to the models for q considered in this study.

Another way of inhancing the models in such a way as to make the parameters more meaningful, would be to put footage drilled into the units

of wildcat footage drilled per cubic mile of oil bearing sediments. Similarly, data for discoveries of oil in place could be expressed as cumulative discoveries of oil in place per cubic mile of oil bearing sediments. The parameters found for various regions would probably have great value to interpret differences between various geological basins.

The data for the above measures were not on hand for this study, and so could not be incorporated into any model. Outside the United States and Canada, such figures would probably not be available. However, it was planned to try some models as described above with the data for Alberta, British Columbia, Saskatchewan and Manitoba in a separate study. Some of these models could probably be improved by appreciating discovery data (classified by year of discovery) to their ultimate values, as indicated in Chapter III, but this was beyond the scope of the present study.

Meanwhile, the value for R obtained by using production figures only in model M16 was consistent with the results obtained from Case III and IV. When production data for the years 1914 to 1968 were used to determine the parameters, and production for the years prior to 1914 was estimated based on these parameters, the calculated values approximated the actual values reasonably well (except that annual production approached but never reached zero in this model.) The calculated and actual values are shown in Table 6.1.6 for odd years only back to 1857.

In order to calculate the parameters necessary for a projection of future production, a slight modification of model M16 to M16S was necessary, as follows. Cumulative production at the end of year t simply became cumulative production for the beginning of year t+1. Using data from 1913 to 1967 for cumulative production, and 1914 to 1968 for annual production in model M16S led to the results shown in Appendix M. By using the parameters thus estimated from model M16S, production could be projected from the year 1968 onward. This projection is shown in Table 6.1.7 at five year intervals up to the year 2020, and at twenty year intervals after that. Maximum production

continued on p.160

TABLE 6.1.6

ESTIMATED AND ACTUAL ANNUAL PRODUCTION FROM 1857 TO 1913
IN THE UNITED STATES, USING MODEL M16 AND DATA
FOR THE YEARS 1914 TO 1968

	ANNUAL PRODUCTION			ANNUAL PRODUCTION	
Year	Estimates	Actual	Year	Estimated	Actual
1857	0.44	0	1887	22	28.3
1859	0.59	0.002	1889	27	35.2
1861	0.78	2.1	1891	34	54.3
1863	1.0	2.6	1893	42	48.4
1865	1.4	2.5	1895	52	52.9
1867	1.8	3.3	1897	64	60.5
1869	2.3	4.2	1899	78	57.1
1871	3.0	5.2	1901	95	69.4
1873	4.0	9.9	1903	110	100.5
1875	5.1	8.8	1905	140	134.7
1877	6.6	13.4	1907	160	166.1
1879	8.4	19.9	1909	200	183.2
1881	11	27.7	1911	230	220.4
1883	14	23.5	1913	270	248.4
1885	17	21.9	Total (all years)	3069.902	3069.902

Estimated figures have been rounded to two-figure accuracy.

Actual figures are from Appendix B.

TABLE 6.1.7

ESTIMATED PRODUCTION FROM 1970 TO 2080 IN THE UNITED STATES, USING MODEL M16S AND DATA FOR THE YEARS 1913 TO 1968

YEAR	ANNUAL PRODUCTION	CUMULATIVE PRODUCTION
1970	3000	89500
1975	3100	105000
1980	3200	121000
1985	3200	137000
1990	3200	153000
1995	3100	169000
2000	3000	184000
2005	2900	199000
2010	2700	214000
2015	2600	227000
2020	2400	239000
2040	1700	280000
2060	1100	308000
2080	680	325000

Estimates of annual production rounded to two-figure accuracy.

Estimates of cumulative production rounded to three-figure accuracy.

was calculated to occur in the year 1985 (which was seven years prior to the year of the maximum rate of 3300 million barrels annually predicted by C.L. Moore using the Gompertz curve, and seventeen years after the maximum of 2900 million barrels predicted by M. King Hubbert using the logistic curve.[1,2]

The ultimate recoverable resource estimated by model M16S was 352.0 billion barrels of oil if t became very large. Weeks estimate of primary recovery of crude oil and natural gas liquid was 270 billion barrels, and production from secondary recovery was estimated to eventually reach 75% of production from primary recovery for total world production.[3,4] As discussed in Chapter III, Hubbert had estimated the resource base at 170-175 billion barrels, while the value Moore estimated for R in the Gompertz equation was 330 to 381 billion barrels, - depending upon the date the estimate was made during the period 1960 to 1966.

The remarkable property of model M16 is that is describes the pattern of production very well for the years 1914 to 1968 - a period encompassing two world wars and a major depression. Throughout this entire period, a constant effect appeared to be manifest in model M16 - that annual production was jointly dependent upon cumulative production, and the apparent amount of resource remaining.

This feature of model M16, and of M17, suggested the means for generalizing the models. If the value of the ultimate resource of oil, R, is known, the models can be simplified considerably. The results from Case III and IV imply that the ultimate resource of oil in place can readily be estimated using data for only a few (15-20) years in a fairly well explored area such as the United States. Model M16 can then provide an estimate for R, the ultimate amount of oil which may be recovered. If the results from the models for Case II, Case III and M16 are consistent, then the estimated value of R could be substituted in the equation for model M15, and we have

$$\Delta y = k y^{m} (R-y)^{n} \tag{4.2.34}$$

with R known.

This equation can also be stated as

$$\ln \Delta y = \ln k + m \ln y + n \ln (R - y) , \qquad (6.1.24)$$

or $$\ln \Delta y = \beta_0 + \beta_1 \ln y + \beta_2 \ln (R - y) \qquad (6.1.25)$$

If R is given, then the last equation is in the familiar linear multivariate form. It would therefore be possible to include as many other variables V_1, V_2, V_3, ... as may be desired, as follows:

Let $$\ln \Delta y = \beta_0 + \beta_1 \mathrm{fn}(V_1) + \beta_2 \mathrm{fn}(V_2) + \ldots$$

$$+ \beta_{k-1} \ln y + \beta_k \ln (R - y) , \qquad (6.1.26)$$

where $\mathrm{fn}(V_i)$ indicates any trnasformation of the data for V_i (such as $\ln(V_i)$, $\exp(V_i)$, etc.). Estimates of the k+1, unknown regression coefficients β_0, β_1, β_2, ... β_k could readily be obtained; the coefficients $\hat{\beta}_{k-1}$ and $\hat{\beta}_k$ would thus assign a weight to the terms $\ln y$ and $\ln (R - y)$ according to their relative predictive value.

The independent variables which could be included for consideration in this fashion could certainly include those suggested in Section Six of Chapter III. A few of the variables which come to mind include:

a) population,
b) total energy demand,
c) cost of domestic oil,
d) cost of imported oil,
e) restrictions on oil imports,
f) national emergency and security considerations,
g) tax incentives,
h) production of other fossil fuels,
i) uranium production and imports,
k) technological developments, such as in secondary recovery,
l) volume of oil bearing rocks found,
m) productive nature of various basins,
and o) implications of North Sea and Artic developments.

The above list was meant to be illustrative only, and not an exhaustive summary of all factors influencing oil production. However, in order to use these variables

in a projection, the values of each variable would have to be estimated for each point in time over which the series of production figures were to be estimates.

Another approach that should probably be tried would be to use one of the models studied to describe the time pattern of the consumption of oil. A reasonable candidate would be model M16S, or a model similar to Equation 6.1.26, but with domestic consumption of all oils as the dependent variable. Such a model could then provide an estimate of the anticipated difference between domestic supply and demand for oil in the United States, and this information would be of great value to an exporting area such as Western Canada.

On the basis of the foregoing discussion, a number of models of the availability and consumption of oil could be postulated. It was hoped that the present study may serve as a basis for such models, and provide the framework against which they could be tested.

REFERENCES:

1. C.L. Moore, Projections of U.S. Petroleum Supply to 1980, United States Department of the Interior, Washington, D.C., December, 1966, p9.

2. M. King Hubbert, Energy Resources, Publication 1000-D, National Research Council, Washington, D.C., 1962, p.63.

3. Lewis G.Weeks, Sixth World Petroleum Congress, Proceedings, Frankfurt, June, 1963, p.271.

4. Gordon H. Barrows, International Petroleum Industry, Volume I, International Petroleum Institute Inc., New York, 1965, p.9.

APPENDIX A

CRUDE OIL PRODUCTION, PROVED RESERVES, AND 1968 ESTIMATE OF ORIGINAL OIL IN PLACE BY YEAR OF DISCOVERY IN THE UNITED STATES EXCLUDING ALASKA

1920 - 1968

Millions of Barrels

I	II	III	IV
YEAR	ANNUAL PRODUCTION	PROVED RESERVES*	ORIGINAL OIL IN PLACE BY YEAR OF DISCOVERY
Pre- 1920	4987 c		85273 a
1920	445 a	7200 b	8627
1921	474	7800	11653
1922	559	7600	4442
1923	733	7600	3333
1924	713	7500	3080
1925	764	8500	4044
1926	771	8800	13159
1927	902	10500	5457
1928	901	11000	7897
1929	1008	13200	9880
1930	897	13600	13655
1931	849	13000	6538
1932	784	12300	2413
1933	905	12000	4117
1934	906	12177	7774
1935	992	12400	8271
1936	1087	13063	19117
1937	1265	15507	8941
1938	1202	17348	10471
1939	1253	18483	5421
1940	1335	19025	8776
1941	1389	19589	6494
1942	1374	20083	4896
1943	1493	20064	4453
1944	1669	20453 b	7582

* Figures for proved reserves from 1920 to 1944 include condensate.

I	II	III	IV
YEAR	ANNUAL PRODUCTION	PROVED RESERVES*	ORIGINAL OIL IN PLACE BY YEAR OF DISCOVERY
1945	1705	19942 a	6539
1946	1728	20874	4943
1947	1851	21488	5749
1948	2003	23280	9108
1949	1825	24649	14405
1950	1952	25268	7160
1951	2213	27468	5391
1952	2258	27961	4854
1953	2315	28945	7546
1954	2274	29561	6380
1955	2427	30012	5196
1956	2560	30435	5592
1957	2560	30300	5802
1958	2384	30536	3357
1959	2495	31716	2602
1960	2473	31586	2615
1961	2503	31674	1706
1962	2543	31305	2528
1963	2602	30895	1543
1964	2634	30908	2494
1965	2688	31193	1781
1966	2850	31131	1177
1967	3017	30996	1542
1968	3058 a	30334 a	784 a
Total....	86575		386558

REFERENCES:

Columns II , III , and IV:

(a) American Petroleum Institute, Reserves of Crude Oil, Natural Gas Liquids, and Natural Gas in the United States and Canada as of December 31, 1968, Volume 23, Washington, D.C., p. 29 and p. 31.

(b) DeGolyer and MacNaughton, Twentith Century Petroleum Statistics, Dallas, Texas, 1968, p. 17.

(c) See Appendix B.

APPENDIX B

CRUDE OIL PRODUCTION AND ESTIMATED CUMULATIVE DISCOVERIES IN THE UNITED STATES,

EXCLUDING ALASKA 1859 - 1919

Millions of Barrels

I YEAR	II ANNUAL PRODUCTION	III ESTIMATED CUMULATIVE DISCOVERIES
1859	0.002 a	
1860	0.5	152 a
1861	2.1	152
1862	3.1	152
1863	2.6	153
1864	2.1	166
1865	2.5	224
1866	3.6	225
1867	3.3	231
1868	3.6	246
1869	4.2	286
1870	5.3	336
1871	5.2	749
1872	6.3	750
1873	9.9	750
1874	10.9	766
1875	8.8	790
1876	9.1	804
1877	13.4	821
1878	15.4	878
1879	19.9	1364
1880	26.3	1365
1881	27.7	1365
1882	30.4	1400
1883	23.5	1403
1884	24.2	1420
1885	21.9	1778
1886	28.1	1880
1887	28.3	1973
1888	27.6	1983
1889	35.2	2006
1890	45.8	2038
1891	54.3	2077
1892	50.5	2139

I	II	III
YEAR	ANNUAL PRODUCTION	ESTIMATED CUMULATIVE DISCOVERIES
1894	49.3	2574
1895	52.9	2613
1896	61.0	2650
1897	60.5	2751
1898	55.4	2755 a
1899	57.1	3146 b
1900	63.6	3304
1901	69.4	3915
1902	88.8	4021
1903	100.5	4441
1904	117.1	4649
1905	134.7	5014
1906	126.5	5281
1907	166.1	5340
1908	178.5	5762
1909	183.2	6181
1910	209.6	6305
1911	220.4	6847
1912	222.9	7401
1913	248.4	7647
1914	265.8	7799
1915	281.1	8227
1916	300.8	8778
1917	335.3	9279
1918	355.9	9606
1919	378.4 a	9944 b
Total	4987.2 b	

REFERENCES:

Column II

(a) World Oil, Vol CXLV, August 15, 1957, p. 193.

(b) Total for the Column.

Column III

(a) C. L. Moore, Methods for Evaluating U.S. Crude Oil Resources and Projecting Domestic Crude Oil Availability, United States Department of the Interior, Office of Oil and Gas, 1962, pp. 47, 94 and 99. (The recovery rate prior to 1930 was 15%, and the 1945 recovery factor was 20.417%, as estimated by Moore. Cumulative discoveries have thus been adjusted by a factor of (15 ÷ 20.417), to convert to recovery conditions as of the year the discoveries were reported.)

(b) Ibid, p. 100.

APPENDIX C

TOTAL WELL COMPLETIONS IN THE UNITED STATES, EXCLUDING ALASKA

1859 - 1968

I. YEAR	II. ANNUAL WELL COMPLETIONS
1859	2 a
1860	240
1861	247
1862	385
1863	595
1864	1,132
1865	1,277
1866	1,142
1867	1,256
1868	1,064
1869	1,682
1870	1,639
1871	1,452
1872	1,166
1873	1,244
1874	1,304
1875	2,458
1876	2,960
1877	3,956
1878	2.988
1879	2,817
1880	4,216
1881	3,876
1882	3,271
1883	3,018
1884	2,365
1885	2,955
1886	4,308
1887	2,142
1888	2,286
1889	6,697
1890	8,831
1891	5,207
1892	3,878
1893	4,276

YEAR	ANNUAL WELL COMPLETIONS	
1894	7,649	
1895	13,250	
1896	14,017	
1897	9,815	
1898	8,897	
1899	14,353	
1900	17,077	
1901	15,097	
1902	16,474	
1903	19,169	
1904	20,545	
1905	16,566	
1906	18,830	
1907	19,668	
1908	16,909	
1909	18,327	
1910	14,940	
1911	13,768	
1912	17,180	
1913	25,590	
1914	23,137	
1915	14,157	
1916	24,619	
1917	23,407	a
1918	25,696	b
1919	29,383	
1920	34,081	
1921	21,985	
1922	24,658	
1923	24,438	
1924	21,894	
1925	25,623	
1926	29,301	
1927	24,141	
1928	22,331	
1929	26,356	
1930	21,240	
1931	12,432	
1932	15,021	
1933	12,312	
1934	18,197	
1935	21,420	

YEAR	ANNUAL WELL COMPLETIONS
1936	25,888
1937	32,560
1938	27,149
1939	27,340
1940	30,040
1941	32,140
1942	18,150
1943	19,477
1944	25,202
1945	26,879
1946	29,228
1947	33,122
1948	39,778
1949	39,038
1950	43,279
1951	44,516
1952	45,821
1953	49,279
1954	53,930
1955	56,682
1956	58,160
1957	53,838
1958	49,107
1959	51,747
1960	46,728
1961	46,914
1962	46,139
1963	43,628
1964	45,215
1965	41,404
1966	37,831
1967	33,744
1968	32,807 b
TOTAL.....	2,189,042

a) American Petroleum Institute, Personal communication, Washington, D.C., January 22, 1970.

b) DeGolyer and MacNaughton, Twentieth Century Petroleum Statistics, 1969, DeGolyer and MacNaughton, Dallas, Texas, 1969, p. 31.

APPENDIX D

FOOTAGE DRILLED, NEW WELLS, IN THE UNITED STATES, EXCLUDING ALASKA 1934 - 1968

I.

YEAR	ANNUAL WELL FOOTAGE (Thousands of Feet)
1934	49,804
1935	61,523
1936	77,072
1937	99,923
1938	87,497
1939	83,855
1940	91,870
1941	95,387
1942	64,813
1943	58,504
1944	80,838
1945	92,982
1946	101,125
1947	112,816
1948	136,709
1949	138,617
1950	159,761
1951	176,757
1952	188,393
1953	198,432
1954	211,296
1955	228,530
1956	235,386
1957	220,864
1958	196,507
1959	206,057
1960	194,402
1961	191,973
1962	197,833
1963	185,729
1964	190,794
1965	178,034
1966	167,612
1967	146,266
1968	148,923
Total	5,056,884

REFERENCE:

DeGolyer and MacNaughton, Twentieth Century Petroleum Statistics, 1969, DeGolyer and

APPENDIX E

DOMESTIC DEMAND FOR ALL REFINED OILS AND FOR GASOLINE IN THE UNITED STATES

1919 - 1968
(Millions of Barrels)

I. YEAR	II. ALL OILS	III. GASOLINE
1919	375	89
1920	457	109
1921	458	117
1922	531	138
1923	652	174
1924	688	197
1925	727	233
1926	781	268
1927	803	305
1928	861	339
1929	941	383
1930	927	398
1931	903	408
1932	835	378
1933	868	380
1934	920	410
1935	984	435
1936	1093	482
1937	1170	519
1938	1137	523
1939	1231	556
1940	1327	589
1941	1486	668
1942	1450	589
1943	1521	568
1944	1671	632
1945	1773	696
1946	1793	735
1947	1990	795
1948	2114	871
1949	2118	914
1950	2375	994
1951	2570	1090
1952	2664	1157
1953	2775	1206
1954	2832	1231

I.	II.	III.
YEAR	ALL OILS	GASOLINE
1955	3088	1334
1956	3213	1373
1957	3219	1393
1958	3315	1436
1959	3450	1485
1960	3536	1512
1961	3579	1533
1962	3735	1583
1963	3854	1635
1964	3959	1658
1965	4125	1702
1966	4325	1793
1967	4481	1843
1968	4788	1925

REFERENCE:

DeGolyer and MacNaughton, Twentieth Century Petroleum Statistics, 1969, DeGolyer and MacNaughton, Dallas, Texas, 1969, p.p. 46-47.

APPENDIX F

CRUDE OIL PRODUCTION IN NORTH AMERICA FOR VARIOUS REGIONS

1862 – 1968

(Thousands of Barrels)

I YEAR	II CANADA	III ALBERTA	IV ALASKA	V MEXICO	VI CUBA
1862	12 a				
1863	83				
1864	90				
1865	110				
1866	175				
1867	190				
1868	200				
1869	220				
1870	250				
1871	270				
1872	308				
1873	365				
1874	169				
1875	220				
1876	312				
1877	312				
1878	312				
1879	575				
1880	350				
1881	275				
1882	275				
1883	250				
1884	250				
1885	250				
1886	584				
1887	526				
1888	695				
1889	705				
1890	745				
1891	755				
1892	780				
1893	798				

I YEAR	II CANADA	III ALBERTA	IV ALASKA	V MEXICO	VI CUBA
1894	829				
1895	726				
1896	727				
1897	710				
1898	758				
1899	808				
1900	913				
1901	757			10 a	
1902	531			40	
1903	487			75	
1904	553			126	
1905	634			251	
1906	569			502	
1907	789			1005	
1908	528			3933	
1909	421			2714	
1910	316			3634	
1911	291			12553	
1912	243			16558	
1913	228			25696	
1914	215			26235	
1915	215			32911	
1916	198			40546	
1917	214 a			55293 a	
1918	305 b			63828 b	
1919	241			87073	
1920	196			157069	
1921	188			193398	
1922	179			182278	
1923	170	2 a		149585	
1924	161	5 b		139678	
1925	332	172		115515	
1926	364	212		90421	
1927	477	326		64121	
1928	624	492		50151	
1929	1117	999		44688	
1930	1522	1445		39530	
1931	1543	1444		33039	
1932	1044	918		32805	
1933	1145	827		34001	23 a
1934	1417	852		38172	28
1935	1447	767		40241	47
1936	1500	718		41028	62

I YEAR	II CANADA	III ALBERTA	IV ALASKA	V MEXICO	VI CUBA
1937	2944	2140		46907	33
1938	6966	6212		38506	78
1939	7838	7297		42898	112
1940	8591	8221		44036	142
1941	10134	9616		42196	150
1942	10365	9834		34815	151
1943	10052	9213		35163	107
1944	10099	8341		38203	109
1945	8483	7643		43547	149
1946	7586	6704		49235	269
1947	7692	6382		56284	300
1948	12287	10505		58508	159
1949	21305	19768		60910	206
1950	29044	27149		72443	156
1951	47615	45836		77312	128
1952	61237	58837		77275	36
1953	80899	76702		72440	17
1954	96080	87593		83653	25
1955	129440	112853		89406	375
1956	171981	143682		90660	543
1957	181848	136766		88266	395
1958	165496	112471		93533	344
1959	184778	128802		96393	192
1960	189534	130499		99049	108
1961	220861	157766	6327 a	106784	80
1962	244139	165098	10259	111830	90
1963	258435	168670	10740	114867	72
1964	274626	175089	11059	115576	264
1965	292308	183729	11128	117959	264
1966	320467	202502	14358	121149	260
1967	351287	230517	29126	133042	756
1968	417691 b	250675 b	66203 a	141960 b	146 a
TOTAL	3880151	2716291	159200	4313508	6376

REFERENCES:

Column II:

(a) G. H. Barrows, International Petroleum Industry, Volume 1, International Petroleum Institute, Inc., New York, 1965, p. 270.

(b) De Golyer and MacNaughton, Editors, Twentieth Century Petroleum Statistics, 1969 (Annual), De Golyer and MacNaughton, Dallas, Texas, 1969, p. 5.

REFERENCES cont'd

Column III:

(a) Petroleum and Natural Gas Conservation Board, Alberta Oil and Gas Industry, Calgary, Alberta, 1952. (Cumulative production up to and including 1922 was given at 63,000 barrels.)

(b) Oil and Gas Conservation Board, Cumulative Annual Statistics, Report 69 - 17, Oil and Gas Conservation Board, Calgary, Alberta, 1968, p. 305.

Column IV:

(a) De Golyer and MacNaughton, p. 23. (Data for Alaska was included with Nevada prior to 1961.)

Column V:

(a) G. H. Barrows, p. 274.

(b) De Golyer and MacNaughton, p. 5.

Column VI:

(a) De Golyer and MacNaughton, p. 5.

APPENDIX G

WORLD CRUDE OIL PRODUCTION

1857 - 1968

(Millions of Barrels)

I YEAR	II NORTH AMERICA	III SOUTH AMERICA	IV TOTAL WORLD
1857			.002 a
1858			.004
1859	.002 a		.006
1860	.500		.509
1861	2.114		2.131
1862	3.069		3.092
1863	2.694		2.763
1864	2.206		2.304
1865	2.608		2.716
1866	3.773		3.899
1867	3.537		3.709
1868	3.846		3.990
1869	4.435		4.696
1870	5.511		5.799
1871	5.475		5.730
1872	6.601		6.877
1873	10.260		10.840
1874	11.100		11.930
1875	9.008		9.980
1876	9.445		11.050
1877	13.660		15.750
1878	15.710		18.420
1879	20.490		23.600
1880	26.640		30.020
1881	27.940		31.990
1882	30.630		35.700
1883	23.700		30.260
1884	24.470		35.970
1885	22.110		36.770

I YEAR	II NORTH AMERICA	III SOUTH AMERICA	IV TOTAL WORLD
1886	28.650		47.240
1887	28.810		47.810
1888	28.310		52.170
1889	35.870		61.510
1890	46.570		76.630
1891	55.050		91.100
1892	51.300		88.740
1893	49.230		92.040
1894	50.170		89.340
1895	53.620		103.700
1896	61.690	.047 a	114.200
1897	61.190	.071	122.000
1898	56.120	.071	125.000
1899	57.880	.089	131.100
1900	64.530	.274	149.100
1901	70.160	.275	167.400
1902	89.340	.287	181.800
1903	101.000	.278	194.900
1904	117.800	.290	217.900
1905	135.600	.373	215.100
1906	127.600	.531	213.300
1907	167.900	.751	264.000
1908	183.000	.957	285.300
1909	186.300	1.486	298.700
1910	213.500	1.421	327.800
1911	233.300	1.763	344.400
1912	239.700	2.236	352.400
1913	274.400	2.706	385.300
1914	292.200	2.757	407.500
1915	314.200	3.842	432.000
1916	341.500	4.389	457.500
1917	390.800 a	5.574 a	502.900 a
1918	420.100 b	6.355 b	503.500 b
1919	465.700	6.285	555.900
1920	600.200	7.068	688.900
1921	665.800	9.649	766.000
1922	740.000	13.210	858.900
1923	882.200	16.860	1016.000
1924	853.800	26.660	1014.000
1925	879.600	40.810	1069.000

I YEAR	II NORTH AMERICA	III SOUTH AMERICA	IV TOTAL WORLD
1926	861.700	67.150	1097.000
1927	965.700	102.800	1263.000
1928	952.200	155.500	1325.000
1929	1053.100	190.800	1486.000
1930	939.100	189.500	1410.000
1931	885.700	168.200	1373.000
1932	819.000	167.800	1310.000
1933	940.800	169.100	1442.000
1934	947.700	196.500	1522.000
1935	1038.000	210.800	1654.000
1936	1142.000	221.900	1792.000
1937	1329.000	258.400	2039.000
1938	1260.000	262.900	1988.000
1939	1316.000	284.200	2086.000
1940	1406.000	268.800	2150.000
1941	1455.000	309.100	2221.000
1942	1432.000	220.200	2093.000
1943	1551.000	257.300	2257.000
1944	1726.000	343.400	2592.000
1945	1766.000	406.500	2595.000
1946	1791.000	466.700	2745.000
1947	1921.000	517.600	3022.000
1948	2091.000	574.900	3433.000
1949	1924.000	573.400	3404.000
1950	2075.000	644.100	3803.000
1951	2373.000	726.700	4283.000
1952	2428.000	766.200	4505.000
1953	2510.000	756.300	4798.000
1954	2495.000	809.700	5017.000
1955	2704.000	910.700	5626.000
1956	2880.000	1037.000	6125.000
1957	2887.000	1169.000	6438.000
1958	2708.000	1121.000	6608.000
1959	2856.000	1204.000	7145.000
1960	2864.000	1266.000	7690.000
1961	2949.000	1319.000	8184.000
1962	3032.000	1439.000	8882.000
1963	3126.000	1468.000	9538.000
1964	3177.000	1531.000	10309.000
1965	3259.000	1564.000	11058.000
1966	3470.000	1549.000	12022.000
1967	3701.000	1650.000	12913.000
1968	3889.000 b	1684.000 b	14077.000 b
TOTAL	96,902.224	29,355.515	210,782.577

REFERENCES:

Column II:

(a) G. H. Barrows, International Petroleum Industry, Volume 1, International Petroleum Institute, Inc., New York, 1965, pp. 268 - 278. (Totals for Canada, Mexico, and United States.)

(b) De Golyer and Mac Naughton, Twentieth Century Petroleum Statistics, 1969, De Golyer and Mac Naughton, Dallas, Texas, 1969, p. 4.

Column III:

(a) G. H. Barrows, pp. 268 - 278. (Totals for Argentina, Bolivia, Brazil, Chili, Columbia, Ecuador, Peru, Trinidad, and Venezuela.)

(b) De Golyer and Mac Naughton, p. 4.

Column IV:

(a) G. H. Barrows, p. 268.

(b) De Golyer and Mac Naughton, p. 4.

Note: Most of the data for this Exhibit have been rounded to four figures; extra zeros have been added merely for ease in reading.

APPENDIX H

COMPUTER PROGRAMS TO MINIMIZE THE SUM OF SQUARES BETWEEN THE OBSERVED AND CALCULATED VALUES

M is a tx2 matrix with the observed (dependent) values z=z(t) in the first column, and observed (predictive) values x=x(t) in the second column. The invoked function FN calculates the values z' = bf(x; u, v,w)=bx which are an estimate for z in the case of regression through the origin.

TEST 3 is an APL\360 computer program which seeks out the values of the parameters u, v and w which minimizes the sum $(z-z')^2$. The accuracy of the parameters can be specified by putting DU, DV and DW in the vector A to be as small as may be desired.

TEST 3 indicates the particular direction of the search(i.e. -- whether u,v and/or w increase or decrease) in minimizing the sum of squares for error. The algorithm then displays the sum of squares for error $\Sigma\,(z-z')^2$ over all the observations, gives some indiation of the relative change in this sum by varying u, v and w, and finally produces a matrix MAT of observed values z in the first column, and calculated values x in the second column. The algorithm is shown below:

```
     ∇ LC←A TEST3 M
[1]    PAR←(U←A[1]),(V←A[2]),(W←A[3]),(DU←A[4]),(DV←A[5]),DW←A[6]
[2]    [illegible]←+/([illegible]-[illegible]←+/[illegible];1])*2
[3]    [illegible]←(27 3)ρ[illegible]←0
[4]    →((J←J+1)>IT)/19
[5]    TE[;1]←U+DU←27ρ 1 0 ¯1 ¯1 ¯1 0 1 1 0
[6]    TE[;2]←V+DV×27ρ ¯1 ¯1 ¯1 0 1 1 1 0 0
[7]    TE[;3]←W+DW×(9ρ1),(9ρ¯1),9ρ0
[8]    TV←27ρI←1
[9]    TV[I]←TE[I;] FN M
[10]   →(I=27)/12
[11]   →9,I←I+1
[12]   [illegible]
[13]   →(R=1)/15
[14]   →4,PA←(U←(TE[R;1]),(V←TE[R;2]),(W←TE[R;3]),R←28-R
[15]   'PARTIAL DERIVATIVES=0,SSE=';TV[27]
[16]   [illegible]←(TV[[illegible]],[illegible],[illegible])-TV[27]
[17]   MAT←⍉(2,NO)ρYO,XO,AT←U,V,W,DU,DV,DW
[18]   →0
[19]   [illegible]←U,V,W,DU,DV,DW
```

If only two parameters (u and v) are to be estimated, TEST 3 can be simplified to the algorithm TEST 2 shown below:

```
        ∇TEST2 [□]∇
      ∇ LS←A TEST2 M
[1]     PAR←(U←A[1]),(V←A[2]),(DU←A[3]),DV←A[4]
[2]     SST←+/(Y←YO-YM←+/YO÷NO←ρYO←M[;1])*2
[3]     TE←(9 2)ρJ←0
[4]     →((J←J+1)>IT)/17
[5]     TE[;1]←U+DU×9ρ 1 0 ¯1 ¯1 ¯1 0 1 1 0
[6]     TE[;2]←V+DV×9ρ 1 1 1 0 ¯1 ¯1 ¯1 0 0
[7]     TV←9ρI←1
[8]     TV[I]←TE[I;] FN M
[9]     →(I=9)/11
[10]    →8,I←I+1
[11]    □←R←(⍋φTV)[1]
[12]    →(R=1)/14
[13]    →4,PAR←(U←TE[R;1]),(V←TE[R;2]),R←10-R
[14]    'PARTIAL DERIVATIVES=0,SSE=';TV[9]
[15]    MAT←⍉(2,NO)ρYO,XO,AP←U,V,DU,DV
[16]    →0
[17]    [illegible]
      ∇
```

Similarly, if only one parameter (u) is to be estimed, TEST 1 as shown below can be used to estimate u:

```
        ∇TEST1 [□]∇
      ∇ LS←A TEST1 M
[1]     PAR←(U←A[1]),DU←A[2]
[2]     SST←+/(Y←YO-YM←+/YO÷NO←ρYO←M[;1])*2
[3]     O←U FN M
[4]     E←(U+DU) FN M
[5]     W←(U-DU) FN M
[6]     →10-(⍋O,E,W)[1]
[7]     →4,VAR←(U←U-DU),O←W
[8]     →4,VAR←(U←U+DU),O←E
[9]     'PARTIAL DERIVATIVE=0,SSE=';O
[10]    MAT←⍉(2,NO)ρYO,XO

      ∇
```

To quickly obtain an initial rough approximation of the parameters u, v and w, the number of observations in M can be reduced considerably, and the number of rows in the matrix TE of TESTS can be reduced from 27 to 9, as illustrated in the algorithm EST3 below:

```
      ∇EST3 [□]∇
    ∇ LS←A EST3 M
[1]   PAR←(U←A[1]),(V←A[2]),(W←A[3]),(DU←A[4]),(DV←A[5]),(DW←A[6]
[2]   SST←+/(Y←YO-YM←+/YO÷NO←ρYO←M[;1])*2
[3]   TE←(9 3)ρJ←0
[4]   →((J←J+1)>IT)/19
[5]   TE[;1]←U+DU×9ρ 1 ¯1 ¯1 1 1 ¯1 ¯1 1 0
[6]   TE[;2]←V+DV×9ρ 1 1 ¯1 ¯1 1 1 ¯1 ¯1 0
[7]   TE[;3]←W+DW×9ρ 1 1 1 1 ¯1 ¯1 ¯1 ¯1 0
[8]   TV←9ρI←1
[9]   TV[I]←TE[I;] EST M
[10]  →(I=9)/12
[11]  →9,I←I+1
[12]  R←(⍋TV)[1]
[13]  →(R=1)/15
[14]  →4,PA←(U←(TE[R;1]),(V←TE[R;2]),(W←TE[R;3]),J←1)-?
[15]  'PARTIAL DERIVATIVES=0,SSE=';TV[9]
[16]  □←TV-TV[9]
[17]  MAT←⍉(2,NO)ρYO,XO,AT←U,V,W,DU,DV,DW
[18]  →0
[19]  □←AT←U,V,W,DU,DV,DW
    ∇
```

In each nonlinear model where regression was carried out through the origin, the second line of the invoked algorithm FN would be the same. For model M10, $\Delta y = -ay \ln (y/R)$, and the function FN used was as follows:

```
∇SSE←R FN M
[1]   X←XO-+/(XO←M[;2]×⍟M[;2]÷R)÷NO
[2]   SSE←+/(YO-XO×B←(+/YO×XO)÷+/XO*2)*2
[3]   ∇
```

For model M13, $\Delta y = cy\,(R-y)$, and line 1 of FN was

```
[1]   X←XO-+/(XO←M[;2]×R-M[;2])÷NO
```

For model M21, $y = bq^c$, and line 1 of FN was

```
[1]   X←XO-(+/XO←M[;2]*R)÷NO
```

For model M22, $y(t) = R(1-e^{c\Psi})$, and line 1 of FN was

```
[1]   X←XO-(+/XO←1-*R×M[;2])÷NO
```

For model M9, $y = R(c^{d^t} - c)$ (so that $y = 0$ when $t = 0$) and line 1 of FN was

```
[1]   X←XO-(+/XO←(R[1]*R[2]*M[;2])-R[1])÷NO
```

For model M12, $y = R(1/(1 + Ab^t) - 1/(1+A))$, (so that $y = 0$ when $t = 0$) and line 1 of FN was

```
[1]   X←XO-(+/XO←(1÷1+R[1]×R[2]*M[;2])-(1÷1+R[1]))÷NO
```

For model M16, $\Delta y = ky\,(1-(y/R)^{\beta})$, and line 1 of FN was

```
[1]   X←XO-(+/XO←M[;2]×1-(M[;2]÷R[1])*R[2])÷NO
```

For model M18, $\Delta y = ky^m (R-y)^{1-m}$, and line 1 of FN was

```
[1]   X←XO-(+/XO←(M[;2]*R[1])×(R[2]-M[;2])*1-R[1])÷NO
```

For model M15, $y = R(1/(1+Ae^{-at})^\theta - 1/(1+A)^\theta)$, (so that $y = 0$ when $t = 0$) and line 1 of FN was

```
[1]   X←XO-(+/XO←((1÷1+R[1]××-R[2]×M[;2])*R[3])
                  -(1÷1+R[1])*R[3]))÷NO
```

For model M17, $\Delta y = ky^m (R-y)^{0.9999}$, and line 1 of FN was

```
[1]   X←XO-(+/XO←M[;2]*R[1])×(R[2]-M[;2])*0.9999)÷NO
```

For the case of a regression line through $\beta_0 \neq 0$, the function FN can be used to calculate the values $z' = b_0 + b_1 \ f(x;u, v, w)$ as an estimate for z. In

```
[2]   SSE←+/(Y-X×B1←(+/X×Y)÷+/X*2)*2
```

and SSE would be the sum of squares for error as obtained in ordinary linear regression. Here, the estimated regression coefficients b_0 and b_1 are given by the normal equations if the parameters above have been estimated, and the sum of squares for error, SSE, can be written as

$$SSE = \Sigma (Y_i - b_0 - b_1 X_i)^2$$

$$= \Sigma (y_i - b_1 x_i)^2$$

with $y_i = Y_i - \bar{Y}$, and $x_i = X_i - \bar{X}$

When using a regression line through $\beta_0 \neq 0$ in models M9*, M12* and M15*, the assumption that $y = 0$ when $t = 0$ is not necessary. Thus, for model M9*, $y = \beta_0 + Rc^{d^t}$, and line 1 of FN was

```
[1]   X←XO-(+/XO←R[1]*R[2]*M[;2])÷NO
```

For model M12*, $y = \beta_0 + R/(1 + Ab^t)$, and line 1 of FN was

```
[1]   X←XO-(+/XO←1÷1+R[1]×R[2]*M[;2])÷NO
```

For model M15*, $y = \beta_0 + R/(1 + Ae^{-at})^\theta$ and line 1 of FN was

```
[1]   X←XO-(+/XO←(1+R[1]×-*R[2]×M[;2])*-R[3])÷NO
```

For some models with a regression line such that β_0 was arbitrary, the first line of the function FN could be simplified because it was not necessary for y to be zero for $t = 0$. For example, in model M22*, line 1 of FN was

```
[1]   X←XO-(+/XO←*R×M[;2])÷NO
```

This was the algorithm used for the models given in Case II, Case III, Case IV, Case V and Case VI. In the model for Case I, FN was the same as for model M22.

REFERENCE:

A.D. Falkoff and K.E. Iverson, APL 360: User's Manual, International Business Machines Corporation, New York, N.Y., 1968.

APPENDIX I

COMPUTER PROGRAM TO CALCULATE STATISTICS FOR REGRESSION THROUGH THE ORIGIN

If Y is a vector of N observations of the dependent variable, and X is a vector of N observations of the independent variable, the program Y REGØ X calculates:

1. The mean of X, $MX = SX \div N$, with $SX = \Sigma X$.
2. The mean of Y, $MY = SY \div N$, with $SY = \Sigma Y$.
3. The sums of squares of X and Y:

$$SX2 = \Sigma X^2, \quad SY2 = \Sigma Y^2,$$

$$MX2 = \Sigma (X-MX)^2, \text{ and } MY2 = \Sigma (Y-MY)^2$$

4. The sums of products:

$$SXY = \Sigma XY,$$

and

$$MXY = \Sigma (X-MX)(Y-MY).$$

5. For the regression equation $Y = \beta_0 + \beta_1 X + \epsilon$, the following statistics:

Estimates b_1 and b_0 of the regression coefficients β_1 and β_0, with $B_1 = b_1 = \frac{\Sigma (X-MX)(Y-MY)}{\Sigma (X-MX)^2}$ as an estimate of β_1, and $BØ = b_0 = MY - B1 \times MX$, as an estimate of β_0.

The sum of squares for error, SSE, with $SSE = \Sigma (Y-Y')^2$, where $Y' = b_0 + b_1 X$.

The sum of squares for regression SSR, with $SSR = (\Sigma XY)^2 \div \Sigma (X-MX)^2$.

The total sum of squares, SST, with $SST = \Sigma (Y-MY)^2$.

The correlation coefficient, r with $r^2 = SSR \div SST$.

6. For the regression equation through the origin, $Y = \beta X + \epsilon$, the following statistics:

An estimate, b, of the regression coefficient β, with $B = b = \Sigma XY \div \Sigma X^2$.
The sum of squares for error, SE, with $SE = \Sigma(Y-Y'')^2$, where $Y'' = bX$.
The standard error of the estimate, EE, with $EE^2 = SE \div (N-1)$.
The F statistic for the hypothesis that $b = 0$, with $F = \Sigma X^2 \div EE$ having 1 and N-1 degrees of freedom.
The F statistic for the hypothesis that $b = MY$, and that $X = 1$,

$$\text{with } F = \frac{(\Sigma XY - \Sigma Y(\Sigma X^2/N)^{1/2})^2}{2(\Sigma X^2 - \Sigma X(\Sigma X^2/N)^{1/2}) \times SSE/(N-2)}$$

7. The algorithm then prepares an Nx4 matrix REM, with the first column being t = 1, 2, 3, ... N, the second column the observed values Y, the third column the calculated values $Y' = bX$, and the fourth column the residuals $(Y-Y')$.

The algorithm REGØ is shown below:[1,2]

```
      ∇REGO [⎕]∇
    ∇ Z←Y REGO X;MX;SX;N;MY;SY;SX2;MX2;SY2;MY2;SXY;MXY;B1;B0;SSE;SSR;SST;B;SE;EE;FM
[1]    'MEAN OF X                          ';MX←(SX←+/X)÷N←ρX
[2]    'MEAN OF Y                          ';MY←(SY←+/Y)÷N
[3]    'SUM OF X SQUARED                   ';SX2←+/X*2
[4]    'SUM OF (X-MX) SQUARED              ';MX2←+/(X-MX)*2
[5]    'SUM OF Y SQUARED                   ';SY2←+/Y*2
[6]    'SUM OF (Y-MY) SQUARED              ';MY2←+/(Y-MY)*2
[7]    'SUM OF XY                          ';SXY←+/X×Y
[8]    'SUM OF (X-MX)(Y-MY)                ';MXY←+/(X-MX)×(Y-MY)
[9]    ⎕←'REGRESSION WITH B0≠0:'
[10]   'ESTIMATE OF B1                     ';B1←MXY÷MX2
[11]   'ESTIMATE OF B0                     ';B0←MY-B1×MX
[12]   'SUM OF SQUARES FOR ERROR           ';SSE←+/(Y-(B0+B1×X))*2
[13]   'REGRESSION SUM OF SQUARES          ';SSR←(MXY*2)÷MX2
[14]   'TOTAL SUM OF SQUARES               ';SST←MY2
[15]   'CORRELATION COEFFICIENT,R          ';(SSR÷SST)*0.5
[16]   'R SQUARED                          ';SSR÷SST
[17]   ⎕←'REGRESSION WITH B0=0:'
[18]   'REGRESSION COEFFICIENT,B           ';B←SXY÷SX2
[19]   'SUM OF SQUARES FOR ERROR           ';SE←+/(Y-B×X)*2
[20]   'STD. ERROR OF THE ESTIMATE         ';EE←(SE÷N-1)*0.5
[21]   'STANDARD ERROR OF B                ';EE÷SX2*0.5
[22]   'F STATISTIC FOR B=ZERO             ';(N-1)×(B*2)×SX2÷SE
[23]   '  DEGREES OF FREEDOM               ';1,(N-1)
[24]   FM←((SXY-SY×(SX2÷N)*0.5)*2)÷2×(SX2-SX×(SX2÷N)*0.5)×SSE÷N-2
[25]   'F STATISTIC FOR B=MEAN             ';FM
[26]   '  DEGREES OF FREEDOM               ';1,(N-2)
[27]   'F STATISTIC FOR B0=0               ';(SE-SSE)÷SSE÷N-2
[28]   '  DEGREES OF FREEDOM               ';1,(N-2)
[29]   REM←⍉(4,N)ρ(⍳N),Y,(B×X),(Y-B×X)
    ∇
```

REFERENCES:

1. A. D. Falkoff and K.E. Iverson, APL\360: User's Manual, International Business Machines Corporation, New York, N.Y., 1968.

2. K.W. Smillie, Personal communication, Department of Computing Science, University of Alberta, June, 1970.

APPENDIX J

COMPUTER PROGRAM TO CALCULATE STATISTICS OF THE RESIDUALS

REM is defined as the following matrix:

Column 1: The year $t = 1,2,3,\ldots N$, ($N = 110$ in models where data were used from 1859 to 1968.)

Column 2: The observed values of the dependent variable, z. (i.e.--The values for the left-hand side of the equation in any model.)

Column 3: a) For the case of regression through the origin, the values estimated for $z' = bx$ from the regression equation $z = bx + e$, where b is the regression coefficient and $e = e(t)$ are the error terms, or residuals.

b) For the case of ordinary linear regression, the values estimated for $z' = b_0 + b_1 x$ from the regression equation $z = b_0 + b_1 x + e$, where b_0 and b_1 are the regression coefficients and $e = e(t)$ are the error terms, or residuals.

In the nonlinear models, the computer algorithm TEST was used to first estimate the values of the parameters u, v and w such that $\sum e^2$ was a minimum.

Column 4: The error terms, or residuals, $e = z-z'$.

In the case of ordinary linear regression, REM was prepared using

the the functions REG, INV, and RES, as described for the University of Alberta APL\360 system. In the case of regression through the origin, REM was prepared using the computer algorithm REGØ described in Appendix I.

Using REM as a basis for all calculations, the main APL 360 function DOC calculates:

I By means of the invoked function DSTAT:

1. The number of residuals, N.
2. The maximum residual.
3. The minimum residual.
4. The range of residuals.
5. The mean of the residuals.
6. The variance of the residuals.
7. The standard deviation of the residuals.
8. The mean deviation of the residuals.
9. The median residual.

II The year of occurence of the residual with maximum absolute value.

III By means of the invoked function SKKU:

1. The coefficient of skewness of the residuals, SK, an estimate of its standard deviation, SDS, and the t value, SK ÷ SDS.

2. The coefficient of kurtosis of the residuals, KU, an estimate of its standard deviation, SDK, and the t value, KU ÷ SDK.

IV An estimate of the serial correlation of the sequence of residuals R(1), R(2), R(3), ..., R(N). The sequence is divided into two series S1 = R(1), R(2), R(3), ..., R(N-1), and S2 = R(2), R(3), R(4), ..., R(N); the simple regression coefficient between S1 and S2 is then calculated.

V By means of the invoked function STATRES:

1. The sum of the N residuals.
2. The sum of squares of the residuals.
3. The Durbin-Watson statistic.
4. The number of positive residuals, p.
5. The number of negative residuals, n.
6. The observed number of runs (m) of consecutively positive or consecutively negative residuals.
7. An estimate of the mean of the population of runs, μ, expected in a random sequence of p positive signs and n negative signs.
8. An estimate of the standard deviation of the population of runs, SD = σ.
9. The t value, i.e. - (MEAN-RUNS) ÷ SD, as calculated above.

VI An estimate of any parabolic trend in time for the residuals; in the equation $R(t) = \beta_0 + \beta_1 t + \beta_1 t^2$, the algorithm estimates:

1. The unknown regression coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.
2. The standard error of $\hat{\beta}_1$ and $\hat{\beta}_2$.
3. The t value of $\hat{\beta}_1$ and $\hat{\beta}_2$.

4. The sum of squares for regression, and the sum of squares for error.
5. The mean square for regression, and the mean square for error.
6. The F value for regression.
7. The standard error of the estimate.
8. The square of the multiple correlation coefficient.

(The invoked fuction REG was used for the above estimates; the F value for regression and the t value for $\hat{\beta}_2$ were the main test statistics desired for this case.)

The functions DOC and SKKU are displayed below. The functions DSTAT, STATRES, CM, REG, INV and RES are a standard part of the University of Alberta APL\360 system.[1]

```
      ∇DOC[⎕]∇
    ∇ ANSWER←DOC REM;R;YM;N;SC;PR
[1]   DSTAT R←REM[;4]
[2]   YM←1968-(N←ρR)-(|R)ι(⌈/|R)
[3]   'YEAR OF MAXIMUM DEVIATION    ';YM
[4]   SKKU R
[5]   SC←CM⍉(2,N-1)ρ(1↓R),¯1↓R
[6]   'SERIAL CORRELATION           ';SC[1;2]
[7]   STATRES REM
[8]   PR←(ι3) REG⍉(3,N)ρ(ιN),((ιN)*2),R
[9]   'PARABOLIC REGRESSION OF RESIDUALS WITH TIME';PR
    ∇
      ∇SKKU[⎕]∇
    ∇ SKKU R;NC;NB;NA;N3;N2;N1;N;S;K2;SK;KU;SDS;SDK
[1]   NC←2+NB←2+NA←4+N3←¯1+N2←¯1+N1←¯1+N←ρR
[2]   S←+/[1](R-(+/R)÷N)∘.*1+ι3
[3]   SK←(N×S[2]÷N1×N2)÷(K2←S[1]÷N1)*1.5
[4]   KU←((N×NA×S[3])-(3×N1×S[1]*2))÷N1×N2×N3×K2*2
[5]   SDS←(6×N×N1÷N2×NA×NB)*0.5
[6]   SDK←((24×N×N1*2)÷N3×N2×NB×NC)*0.5
[7]   'COEFFICIENT OF SKEWNESS      ';SK
[8]   '      STANDARD DEVIATION     ';SDS
[9]   '      T-VALUE                ';SK÷SDS
[10]  'COEFFICIENT OF KURTOSIS      ';KU
[11]  '      STANDARD DEVIATION     ';SDK
[12]  '      T-VALUE                ';KU÷SDK
    ∇
```

REFERENCE:

1. K.W. Smillie, STATPACK2: AN APL STATISTICAL PACKAGE, Second Edition, Publication No. 17, Department of Computing Science, The University of Alberta, Edmonton, February 1969, pp. 16, 26, 28, 29, 32, 40.

APPENDIX K

RESULTS FROM THE NONLINEAR MODELS M10*, M13*, M21*, M22*, M9*, M12*, M16*, M18*, M15* AND M17*, OBTAINED BY USING A REGRESSION LINE THROUGH $\beta_0 \neq 0$.

SUMMARY OF RESULTS FROM THE NONLINEAR MODELS M10*, M13*, M21* AND M22*

STATISTIC	M10*	M13*	M21*	M22*
Number of observations	110	110	109	110
Degrees of freedom for error	107	107	106	107
Error sum of squares	620800	806000	7.218E7	6.843E8
Standard error of the residuals	76.17	86.79	825.2	2529
Correlation, r^2	0.9937**	0.9918**	0.9988**	0.9889*
Estimates of the parameters, inclduing $\hat{\beta}_0$:				
Variable parameter, u	522600	170700	1.239	7.679E-7
Regression coefficient, b_0	-20.94	36.21	110.7	-1809
Regression coefficient, b_1	-0.01917	3.782E-7	0.04391	-20770
$t_1 = b_1 \div$ standard error	-130.1**	114.1**	302.0**	-98.26**
Dependent variable	Δy	Δy	y	y

** Significant at the 0.1% level.

APPENDIX K (continued)

STATISTICS OF THE RESIDUALS IN THE NONLINEAR MODELS M10*, M13*, M21* AND M22*

STATISTIC	M10*	M13*	M21*	M22*
Number of residuals	110	110	109	110
Maximum deviation	213.3	267.1	2922	4927
Year of maximum	1956	1968	1968	1952
Range of deviations	418.7	430.7	4931	9511
Mean deviation	53.90	61.20	550.9	2114
Median residual	-1.290	-26.88	-130.3	615.1
Coefficient of skewness	0.3007	1.128	0.4208	-0.1110
t-value for skewness	1.305	4.894**	1.818	-0.4818
Coefficient of kurtosis	0.7852	1.844	2.055	-0.7660
t-value for kurtosis	1.718	4.035**	4.476**	-1.676
Dependent variable	Δy	Δy	y	y

** Significant at the 0.1% level.

APPENDIX K (continued)

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE NONLINEAR MODELS M10*, M13*, M21*, AND M22*

TEST STATISTICS	M10*	M13*	M21*	M22*
Durbin-Watson statistic	0.6031	0.4883	0.1890	0.01553
Number of runs of positive and negative signs:				
Observed runs, m	15	16	12	4
Positive signs, p	54	39	36	61
Negative signs, n	56	71	73	49
Population mean, μ	55.98	51.35	49.22	55.35
Standard deviation, σ	5.218	4.774	4.592	5.157
t-value, $(\mu - m)/\sigma$	7.854**	7.404**	8.106**	9.956**
Parabolic regression of the residuals with time:				
t-value of $\hat{\beta}_2$	1.290	-2.479	-1.108	5.700**
F for regression	1.033	4.579	1.324	17.96**

** Significant at the 0.1% level.

APPENDEX K (continued)

SUMMARY OF RESULTS FROM THE NONLINEAR MODELS M9*, M12*, M16* AND M18*

STATISTIC	M9*	M12*	M16*	M18*
Number of observations	110	110	110	110
Degrees of freedom for error	106	106	106	106
Error sum of squares	6.902E6	1.535E7	5.948E5	6.275E5
Standard error of the residuals	255.2	380.5	74.91	76.94
Correlation, r^2	0.9999**	0.9998**	0.9939**	0.9936**
Estimates of the parameters, including $\hat{\beta}_0$:				
Variable parameter, u	3.448E-7	1742	333000	0.7574
Variable parameter, v	0.9812	0.9315	0.1994	123300
Regression coefficient, b_o	164.1	-420.0	-3.802	-26.04
Regression coefficient, b_1	546400	147300	0.1432	0.04153
$t = b_1 \div$ standard error	983.8**	659.6**	132.9**	129.4**
Dependent variable	y	y	Δy	Δy

** Significant at the 0.1% level.

APPENDIX K (continued)

STATISTICS OF THE RESIDUALS IN THE NONLINEAR MODELS M9*, M12*M16* AND M18*

STATISTIC	M9*	M12*	M16*	M18*
Number of residuals	110	110	110	110
Maximum deviation	662.9	962.4	192.7	-185.8
Year of maximum	1957	1968	1956	1932
Range of deviations	1141	1689	384.6	370.6
Mean deviation	198.7	303.8	48.65	55.49
Median residual	-16.85	44.19	0.1312	-4.309
Coefficient of skewness	0.5623	-0.001794	0.4076	0.3523
t-value for skewness	2.440	-0.007783	1.769	1.529
Coefficient of kurtosis	0.2199	-0.3656	0.7542	0.4039
t-value for kurtosis	0.4811	-0.7999	1.650	0.8838
Dependent variable	y	y	Δy	Δy

APPENDIX K (continued)

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE NONLINEAR MODELS M9*, M12*, M16* AND M18*

TEST STATISTICS	M9*	M12*	M16*	M18*
Durbin-Watson statistic	0.09285	0.06889	0.6334	0.6093
Number of runs of positive and negative signs:				
Observed runs, m	7	7	17	13
Positive signs, p	53	59	55	52
Negative signs, n	57	51	55	58
Population mean, μ	55.93	55.71	56	55.84
Standard deviation, σ	5.213	5.192	5.220	5.204
t-value, $(\mu - m)/\sigma$	9.386**	9.381**	7.471**	8.231**
Parabolic regression of the residuals with time:				
t-value of $\hat{\beta}_2$	-2.049	3.985**	0.2311	1.229
F for regression	2.431	9.277**	0.02929	0.8842

** Significant at the 0.1% level.

APPENDEX K (continued)

SUMMARY OF RESULTS FROM THE NONLINEAR MODELS M15* AND M17*

STATISTIC	M15*	M17*
Number of observations	110	110
Degrees of freedom for error	105	105
Error sum of squares	6.577E6	597700
Standard error of the residuals	250.3	75.45
Correlation, r^2	0.9999**	0.9939**
Estimates of the parameters, including $\hat{\beta}_0$:		
Variable parameter, u	0.5077	0.8224
Variable parameter, v	0.02059	249000
Variable parameter, w	33.74	0.9999 (maximum)
Regression coefficient, b_0	145.4	-12.53
Regression coefficient, b_1	490600	1.560E-6
$t = b_1 \div$ standard error	1008**	132.6**
Dependent variable	y	Δy

** Significant at the 0.1% level.

APPENDIX K (continued)

STATISTICS OF THE RESIDUALS IN THE NONLINEAR MODELS M15 * AND M17*

STATISTIC	M15*	M17*
Number of residuals	110	110
Maximum deviation	654.7	185.1
Year of maximum	1930	1923
Range of deviations	1118	369.9
Mean deviation	190.4	51.15
Median residual	-16.14	-4.198
Coefficient of skewness	0.6024	0.4461
t-value for skewness	2.614	1.936
Coefficient of kurtosis	0.3851	0.6403
t-value for kurtosis	0.8426	1.401
Dependent variable	y	Δy

APPENDIX K (continued)

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE NONLINEAR MODELS M15* AND M17*

TEST STATISTIC	M15*	M17*
Durbin-Watson statistic	0.09590	0.6337
Number of runs of positive and negative signs:		
Observed runs, m	9	15
Positive signs, p	53	52
Negative signs, n	57	58
Population mean, μ	55.93	55.84
Standard deviation, σ	5.213	5.204
t-value, $(\mu - m)/\sigma$	9.002**	7.847**
Parabolic regression of the residuals with time:		
t-value of $\hat{\beta}_2$	-1.790	0.5747
F for regression	1.845	0.1852

** Significant at the 0.1% level.

APPENDIX L

RESULTS FROM THE NONLINEAR MODELS FOR CASE I, CASE II, CASE III, CASE IV, CASE V, AND CASE VI

SUMMARY OF RESULTS FROM THE MODEL FOR CASE I

STATISTIC	
Number of observations	35
Degrees of freedom for error	33
Estimates of the parameters:	
Variable parameter, c	3.167E-7
Regression coefficient, b	4.427E5
Error sum of squares	3.976E9
Standard error of the estimate	1.098E4
Correlation between observed and calculated values, r^2	0.9699**
Value of F for the hypothesis that $b_o = 0$	1.798
Dependent variable	u

** Significant at the 0.1% level.

APPENDIX L (continued)

SUMMARY OF RESULTS FROM THE MODELS FOR CASE II, CASE III, CASE IV, CASE V AND CASE VI

STATISTIC	CASE II	CASE III	CASE IV	CASE V	CASE VI
Number of observations	35	20	15	35	35
Degrees of freedom for error	32	17	12	32	32
Error sum of squares	1.879E8	1.709E8	4.306E6	5.424E7	5.986E7
Standard error of the residuals	2423	3171	599.0	1302	1368
Correlation between observed and calculated values, r^2	0.9985**	0.9954**	0.9979**	0.9965**	0.9977**
Estimates of the parameters:					
Variable parameter, c	6.410E-7	6.746E-7	5.474E-7	4.074E-8	2.024E-7
Variable parameter, δ	-1.123E6	-1.164E6	-7.329E5	-1.044E6	-1.199E6
Regression coefficient, b_o	3.930E5	3.880E5	3.977E5	3.755E5	1.612E5
Regression coefficient, b_1	-8.072E5	-8.512E5	-5.940E5	-3.919E5	-2.055E5
$t_1 = b_1 \div$ standard error	-148.3**	-62.62**	-78.36**	-97.56	-118.7**
Dependent variable	u	u	u	y	q

** Significant at the 0.1% level.

APPENDIX L (continued)

STATISTICS OF THE RESIDUALS IN THE MODELS FOR CASE I, CASE II, CASE III, CASE IV, CASE V, CASE VI

STATISTIC	CASE I	CASE II	CASE III	CASE IV	CASE V	CASE VI
Number of residuals	35	35	20	15	35	35
Mean deviation	8343	1710	2443	424.0	1102	1071
Maximum deviation	-31270	-5430	5670	1040	3113	3411
Year of maximum	1934	1934	1936	1957	1968	1968
Range of deviations	47230	10710	10240	1900	4676	5579
Mean of residuals	-427.2	0	0	0	0	0
Median residual	668.9	-342.6	-499.6	49.77	-461.2	227.9
Coefficient of skewness	-0.8942	0.1294	0.2663	0.3013	0.6287	0.4073
t-value for skewness	-2.248	0.3254	0.5200	0.5194	1.581	1.024
Coefficient of kurtosis	1.240	0.6264	-0.8346	-0.3855	-0.7056	-0.1243
t-value for kurtosis	1.594	0.8053	-0.8410	-0.3439	-0.9072	-0.1598

APPENDIX L (continued)

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE MODELS FOR CASE I, CASE II, CASE III, CASE IV, CASE V AND CASE VI

TEST STATISTIC	CASE I	CASE II	CASE III	CASE IV	CASE V	CASE VI
Durbin-Watson statistic	n.a.	1.040	1.092	1.079	0.1300	0.2555
Number of runs of positive and negative signs:						
Observed runs, m	3	10	5	7	4	8
Positive signs, p	18	16	8	8	15	19
Negative signs, n	17	19	12	7	20	16
Population mean, μ	18.49	18.37	10.6	8.467	18.14	18.37
Standard deviation, σ	2.912	2.892	2.085	1.857	2.853	2.892
t-value, $(\mu - m)/\sigma$	5.318**	2.895*	2.686	0.7898	4.957**	3.586*
Parabolic regression of the residuals with time:						
t-value of $\hat{\beta}_2$	-17.68**	-0.4294	0.05444	-0.6719	-0.8422	-0.07726
F for regression	164.3**	0.1181	0.004243	0.2553	0.4872	0.08015

* Significant at the 1% level.
** Significant at the 0.1% level.
n.a. Test not applicable.

APPENDIX M

RESULTS FROM THE NONLINEAR MODEL M16, USING DATA FROM 1914 TO 1918

SUMMARY OF RESULTS

STATISTIC	
Number of observations	55
Degrees of freedom for error	52
Estimates of the parameters:	
Variables parameter, R	353900
Variable parameter, p	0.1674
Regression coefficient, b	0.1609
Error sum of squares	584900
Standard error of the estimate	106.1
Correlation between observed and calculated values, r^2	0.9850**
Value of F for the hypothesis that $b_o = 0$	0.03416
Dependent variable	Δy
Independent variable	y

** Significant at the 0.1% level.

APPENDIX M (continued)

STATISTICS OF THE RESIDUALS IN THE NONLINEAR MODEL M16 USING DATA FROM 1914 TO 1968

STATISTIC	ALL YEARS (1914 - 1968)	EVEN YEARS (1914 - 1968)	ODD YEARS (1915 - 1967)
Number of residuals	55	28	27
Mean deviation	88.28	82.64	94.35
Maximum deviation	-196.7	-196.7	178.3
Year of maximum	1932	1932	1923
Range of deviations	393.3	393.3	303.9
Mean of residuals	1.264	-8.385	11.27
Median residual	-22.47	-20.02	-22.47
Coefficient of skewness	0.2124	0.1642	0.2429
t-value	0.6603	0.3727	0.5424
Coefficient of kurtosis	-1.117	-0.7291	-1.528
t-value	-1.764	-0.8494	-1.752

APPENDIX M (continued)

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE NONLINEAR MODEL M16, USING DATA FROM 1914 TO 1968

TEST STATISTIC	ALL YEARS (1914 - 1968)	EVEN YEARS (1914 - 1968)	ODD YEARS (1915 - 1967)
Durbin-Watson statistic	n.a.	n.a.	n.a.
Number of runs of positive and negative signs:			
Observed runs, m	12	10	10
Positive signs, p	26	13	13
Negative signs, n	29	15	14
Population mean, μ	28.42	14.93	14.48
Standard deviation, σ	3.662	2.583	2.544
t-value, $(\mu - m) / \sigma$	4.483**	1.908	1.762
Parabolic regression of the residuals with time:			
t-value of $\hat{\beta}_2$	0.3023	0.4451	0.004711
F for regression	0.06702	0.09956	0.05043

** Significant at the 0.1% level.
n.a. The test is not applicable.

APPENDIX N

RESULTS FROM THE NONLINEAR MODEL M16S, USING DATA FROM 1913 TO 1968

SUMMARY OF RESULTS

STATISTIC	
Number of observations	55
Degrees of freedom for error	52
Estimates of the parameters:	
Variables parameter, R	352000
Variable parameter, p	0.1200
Regression coefficient, b	0.2206
Error sum of squares	631200
Standard error of the estimate	110.2
Correlation between observed and calculated values, r^2	0.9838**
Value of F for the hypothesis that $b_o = 0$	0.03658
Dependent variable	$\Delta y(t)$
Independent variable	$y(t-1)$

** Significant at the 0.1% level.

APPENDIX N (continued)

STATISTICS OF THE RESIDUALS IN THE NONLINEAR MODEL M16S

STATISTIC	ALL YEARS (1914 - 1968)	EVEN YEARS (1914 - 1968)	ODD YEARS (1915 - 1967)
Number of residuals	55	28	27
Mean deviation	91.79	85.81	98.27
Maximum deviation	-207.5	-207.5	189.4
Year of maximum	1932	1932	1923
Range of deviations	409.4	409.4	321.4
Mean of residuals	1.358	-8.779	11.87
Median residual	-25.28	-21.59	-25.28
Coefficient of skewness	0.2149	0.1496	0.2583
t-value for skewness	0.6679	0.3395	0.5768
Coefficient of kurtosis	-1.103	-0.7229	-1.516
t-value for kurtosis	-1.741	-0.8422	-1.738

APPENDIX N (continued)

TESTS OF THE RANDOMNESS OF THE RESIDUALS IN THE NONLINEAR MODEL M16S

TEST STATISTIC	ALL YEARS (1914 - 1968)	EVEN YEARS (1914 - 1968)	ODD YEARS (1915 - 1967)
Durbin-Watson statistic	n.a.	n.a.	n.a.
Number of runs of positive and negative signs:			
Observed runs, m	12	10	10
Positive signs, p	26	13	13
Negative signs, n	29	15	14
Population mean, μ	28.42	14.93	14.48
Standard deviation, σ	3.662	2.583	2.544
t-value, $(\mu - m) / \sigma$	4.483**	1.908	1.762
Parabolic regression of the residuals with time:			
t-value of $\hat{\beta}_2$	0.2907	-0.4365	-0.001566
F for regression	0.05778	0.09710	0.04600

** Significant at the 0.1% level.